DISSERTATION

zur Erlangung des akademischen Grades
Doktor der Informatik (Dr. rer. nat.)

eingereicht an der
Fakultät für angewandte Informatik der Universität Augsburg

von
Simon Flutura

August 2020

Erstgutachter: Elisabeth André

Zweitgutachter: Björn Schuller

Tag der mündlichen Prüfung: Augsburg, 18.6.2021

Zusammenfassung

Smartphones haben Einzug in unseren Alltag erhalten und bilden dort ein zunehmend unentbehrliches Werkzeug. Folglich findet auch die maschinelle Verarbeitung der natürlichen Kommunikation, im Forschungsgebiet der Mensch-Maschine-Interaktion (Social Signal Processing), zunehmend im Alltag statt. Die vorliegende Arbeit betrachtet zum einen das Verarbeiten sozialer Signale mit mobilen Geräten in Alltagssituationen („in the wild"). Dies geschieht etwa am Beispiel von Lacherkennung. Zum anderen wird darüber hinaus aber auch das Wohlbefinden des Nutzers in urbanen Räumen, aufgefasst als lokale Klimazonen, untersucht.

Über den Verlauf dieser Arbeit wird MobileSSI entwickelt, ein Open-Source Framework zur Erkennung sozialer Signale in Echtzeit. Es basiert auf dem etablierten SSI Framework und bietet somit die synchronisierte Aufzeichnung mehrerer Sensoren und darauf basierendes maschinelles Lernen. MobileSSI bringt diese Funktionalität auf mehrere Platformen (Android, Linux, Windows) und erweitert SSI um interaktives maschinelles Lernen, das auch einen Mehrwert im Bezug auf Personalisierung und Privatsphäre bietet. Um MobileSSI in diversen Szenarien im Alltag als Komponente zum Erfassen und Steigern von Wohlbefinden einsetzen zu können, wird Rapidprototyping eingesetzt. Dadurch entstehen nicht nur Konfigurationen der Verarbeitungsabläufe, sondern auch Nutzerschnittstellen.

In einem ersten Anwendungsszenario wurde das vergnügte Zusammensein in der Gruppe als Faktor für Wohlbefinden betrachtet. Dazu wurde der Einsatz mobiler Sensorik zur Lacherkennung - insbesondere Bewegungs- und Audiodaten - als Indikator für Vergnügen einer Gruppe sowie ein asynchrones Fusionsverfahren zur Integration möglicherweise zeitversetzter Lacher mehrerer Personen untersucht. Das Verfahren kam in einem Demonstrator zur Erkennung der Stimmung in einer Gruppe zum Einsatz.

Als zweites Szenario wurde das Monitoring von gesundheitsrelevantem Verhalten im Alltag am Beispiel des Trinkverhaltens von Personen betrachtet. Hierzu wurde Trinkaktivität anhand von mit Smartwatches aufgezeichneten Armbewegungen erkannt. Da es nicht praktikabel ist, die für das maschinelle Lernen erforderlichen Bewegungsdaten im Nachhinein zu annotieren, wurde untersucht, wie sich Annotationsprozesse in den Alltag von Personen integrieren lassen. Durch den Einsatz von interaktivem maschinellen Lernen in Verbindung mit aktivem Lernen wurde zum einen der Arbeitsaufwand von Personen durch die Auswahl besonders relevanter Daten im Lernprozess verringert. Zum anderen wurde eine Nutzungsschnittstelle für eine Smartwatch entwickelt, die eine bequeme Korrektur von Systemvorhersagen im Alltag ermöglicht. Es wurde untersucht, inwieweit der Einsatz des integrierten DrinkWatch Systems im Alltag von Personen angenommen wird. Hierbei kam Bodystorming zum Einsatz, ein gängiges

Verfahren aus dem Bereich Usability Engineering, das die Berücksichtigung körperlicher Erfahrungen betont, um Einsichten zur Entwicklung und zum Einsatz von Technologie zu gewinnen.

In einem dritten Szenario wurde das Potential mobiler Sensorik zur Erfassung möglicher gesundheitsrelevanter Effekte von urbanen Waldstrukturen in Zusammenarbeit mit dem Institut für Geographie der Universität Augsburg erforscht. Konkret wurde untersucht, ob physiologische Signale und Audiodaten Aufschluss über die Umgebung und das Wohlbefinden des Nutzers erlauben. Grundlage bildete eine Modellierung des Klimas anhand von Temperatur und Luftfeuchtigkeit. Der Beitrag dieser Arbeit beinhaltet den Entwurf und die Umsetzung von Technologie und Methoden zum Erfassen und Annotieren von umweltbezogenen Daten in alltäglichen Situationen und wurde von Nutzern bei Begehungen von Routen mit unterschiedlichen Bebaungsformen und Bewuchs validiert.

Schlagwörter:
Mobile Verarbeitung sozialer Signale, Affective Computing, Wohlbefinden, Maschinelles Lernen

Abstract

The ubiquity of smart devices is increasingly shaping our daily lives. Data processing of natural communication with computers, the goal of Social Signal Processing, is also moving beyond controlled settings with the use of mobile computers. Instead of executing data collection in the lab, it is now realized ”in the wild”. This means that data can now be collected, processed and evaluated in everyday situations. The challenges of this thesis lie on the one hand in classical Social Signal Processing, transferred into ”the wild”, by studying laughter recognition. On the other hand challenges go beyond classical Social Signal Processing into affect recognition in relation to urban environments viewed in *local climate zones*.

Throughout this thesis MobileSSI, an open source framework for real-time recognition of social signals is developed. It builds upon the well established SSI framework and thus provides multi-sensor data-recording, and machine-learning capabilities. MobileSSI brings those features to a variety of platforms (Android, Linux, Windows) and extends the capabilities of SSI with interactive machine learning for increased personalization and privacy. Using rapid prototyping in configuration and mobile user interfaces, MobileSSI forms the technical contribution, that is employed throughout different scenarios ”in the wild”, to measure and improve wellbeing.

As a first field of application, group enjoyment was considered as aspect of wellbeing. Mobile sensors were employed to recognize laughter based on accelerometer and audio data. Asynchronous fusion was used to aggregate laughter events also when they occur staggered. The technique led to live demonstration of group enjoyment recognition.

Drink activity as representation of health related behavior in everyday living was used as a second scenario. Smartwatches were used to record and recognize drink activity. Since it is not feasible to annotate motion data, recorded with smartwatches retrospectively, the annotation process has to be adapted in such a way, that it can be executed ”in the wild”. Therefore, Interactive machine learning combined with Active Learning was implemented, to limit the labeling effort to selected data that has the biggest training effect for the machine learning model. Moreover, a user interface for a smart watch was created that allows the comfortable correction of predictions by the system. The evaluation of the system ”in the wild” was realized with bodystorming by users with a prototype. Bodystorming is a common practice in usability engineering with focus on embodied experience, to foster insights for the design and development process of technology.

As a third scenario, mobile sensors were used to measure wellbeing in the context of urban forests in collaboration with the Institute of Geography of the University of Augsburg. In detail, physiological and audio data were analyzed for the recognition of local climate zones and

the users' wellbeiṅg. The study is based on models of urban climate (heat, humidity). The contribution of this thesis includes the design and implementation of techniques and methods to collect and annotate environment-related data "in the wild" that have been validated with users walking along routes comprising varying urban structural types.

Keywords:
Mobile Sociale Signal Processing, Affective Computing, Wellbeing, Machine Learning

Danksagung

”There comes a time in life when everything seems to make sense
… and this is not one of those times” – Adrien Brody

.. and thanks goes to the HCM-Lab for good time and constructive environment.

Join us now and share the software;
You'll be free, hackers, you'll be free.
Join us now and share the software;
You'll be free, hackers, you'll be free.

Hoarders can get piles of money,
That is true, hackers, that is true.
But they cannot help their neighbors;
That's not good, hackers, that's not good.

When we have enough free software
At our call, hackers, at our call,
We'll kick out those dirty licenses
Ever more, hackers, ever more.

Join us now and share the software;
You'll be free, hackers, you'll be free.
Join us now and share the software;
You'll be free, hackers, you'll be free.

Richard Stallman

Contents

List of Figures

List of Tables

Chapter 1.
Introduction

The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.

– Mark Weisser

Mobile computing has become an integral component of everyday living. Maps printed on paper today are a rare sight, and the skill to map oneself's real world position onto the sheet of paper, that represents an area, is slowly disappearing. Even interpersonal interaction gets mediated by smart devices, not just by chatting via instant message services but also by online-services to "match" people and bring them together.

In technological advancement, speech recognizers, such as Mozilla's DeepSpeech[1], are stepping stones to smart watches, that are lacking space, even for virtual keyboards [142]. Audio processing shifts the human computer interaction paradigm from active input by the user to active listening by the computer. At the same time smart devices are brought to situations of natural human interaction "in the wild" that computers in labs rarely witnessed, e.g. being in the pocket while we have dinner together. Subsequently, mobile computing transforms the society, changes the way we communicate and meet. As a consequence it has prospects of changing public organization structure, including the health sector [46], which often is summarized under the label of *M-Health*.

During the COVID-19 crisis the role of those ubiquitous computers can be seen clearly. For one in apps such as "Corona Warn App" [2] that help individuals in judging their risk of an infection.

[1] https://research.mozilla.org/machine-learning/
[2] https://github.com/corona-warn-app

For another, for data donation such as ”Corona-Datenspende”[3] by the Robert Koch Institute, to scientifically examine the spread of the disease by observing physiological data samples. Both apps show the importance of the single citizen's agency regarding his or her data. In case of the warn app, agency lies in collecting as few data as necessary and working decentralized, based on a public discourse, and the pro-active donation of data in the case of the ”Corona-Datenspende” app [191]. While data collection is a topic that comes naturally with mobile technology, when it concerns advertising and social networks as provided by big corporations, data collection is viewed critically when used by public organizations. Smart phones are present when and where a doctor could not. They as well create the opportunity for people to become more self-responsible. At the same time, perceived privacy, of not having to see a doctor, hides data-privacy risks involving corporations and insurances.

There is a wide catalog of risks and chances [3] bound to *M-Health*, the use of mobile-devices in the health-sector. Gathering user-data to process them on centralized servers seems to be an inevitable ingredient to modern machine learning solutions, just as much, as the user seems powerless in the process of creating machine learning models. This lack of power exists since the user's data are collected and processed by corporations, without his involvement, and the result is implemented into systems as intransparently models are created. Furthermore, mobile applications can advance in respect of predictive, preventive, personalized, participatory and psycho-cognitive aspects to tackle obstacles in M-Health adoption [82]. In predicting health risks, behavior change could be motivated in an empathic way, regarding the affective state of the user. To shape the use of mobile devices according to this perception of M-Health, tools and work-flows have to be created, with MobileSSI those are of concern to the text ahead. The research presented in this thesis transfers active sensing as natural interaction in the lab, to mobile sensing ”in the wild”, viewed in relation to wellbeing. As such this thesis can be seen as groundwork for M-Health applications.

Emotion and activity recognition is explored as well as recognition of environmental influences on the body. To identify and integrate the features of complex signals into a recognition process, machine learning is a promising and popular approach in state-of-the-art research. The machine learning process is viewed as a whole within this thesis, from data collection to model evaluation.

Firstly, multi-modal recognition is adapted to the circumstances ”in the wild” in the realization of a mobile laughter recognizer based on auditive cues and chest-movement. Asynchronous fusion is extended to take multiple users into account.

Secondly, the learning process is reshaped to an interactive approach that supports labeling in the wild, where the recording of ground truth video evidence is impracticable. Furthermore,

[3] https://corona-datenspende.de/faq/

interactive machine learning is happening on device and as such is personalized and decentralized to preserve users' privacy.

1.1 Research Objectives

The research objectives of this thesis question the role users and their devices play, in the creation of mobile machine learning models related to wellbeing. Building upon proven technology of Social Signal Interpretation in the lab, a mobile solution, MobileSSI is elaborated over the course of this work. The mobile approach is concerned with affective computing tasks "in the wild". This involves the design of scenarios, spanning different aspects of wellbeing and tackling different challenges of data-processing on mobile devices.

1.1.1 Rapid Application Development for the Wild

Mobile applications have to cope with varying requirements with respect to sensors and user-interfaces. To flexibly design the flow of signal processing, an XML-based definition of processing pipelines is already used on the desktop. To cope with the additional requirement of increased user interaction in label acquisition and visualization, rapid prototyping is extended by the use of Web-based user interfaces. The need for integrating acquisition of data, ground truth and machine learning has been identified in the field of mobile sensing [111]. Custom UI is of increased importance not just due to the increased user involvement in the creation process, but also due to the combination of devices of different form factors. Applications target distributed ensembles of devices – wrist bands, smart watches, smart phones and tablets – rather than a single device with a standardized screen size. Since mobile devices often are used while pursuing a different primary action, those interfaces have to be efficient and tailored to the task. Web-based interfaces are widespread and highly customizable. They offer the possibility to distribute input, processing and output and thus meet the demand for applicability in rapid prototyping.

How can rapid prototyping be extended to involve users in mobile processes of labeling, machine learning and signal processing within different scenarios?

1.1.2 Heterogeneous Ubiquitous Input

Mobile devices are used in diverse environments and contexts, those different scenarios often require a varying combination of sensors. Microphone and accelerometer might fit one task in a certain environment, whereas skin conductance and heart rate are the better fit for another. Sensors help to identify important context – is the user drinking or uncomfortable due to heat

stress – but can be combined in case of noise or interference on one signal. Fusion of multiple signals and sensors has the potential to increase reliability not just by increasing redundancy, but also in making a solution robust in different contexts and situations. Individual sensors and modalities convey information at different speeds. Widely used fusion algorithms operate at the level of feature or classifiers, both locked to operate in sync on a per-frame basis. To tackle the different nature of each signal, individual processing and asynchronous fusion [110] is required. Furthermore, fusion might not just be used between modalities but between different devices and persons as well, at a higher level of abstraction. This higher level of abstraction can bring Mobile Social Signal Processing towards interaction with social behavior, instead of individual behavior cues. This is considered a key area of advancement by Palaghias et al. [143]. Are tailored approaches for both fusion and synchronization, tested in stationary environments, of value in a mobile application context?

1.1.3 On Device Machine Learning

Machine learning on data collected using mobile devices often conflict with users' privacy. Cloud based services are considered a major challenge in mobile emotion sensing [198]. By relying on technology that runs machine-learning processes locally on smart devices and adapting them for interactive use, critical data has not to be collected centrally or even collected at all. The involvement of users in the machine learning process is described as interactive machine learning (iML). Ideally users label their own data and approve the models quality in an evaluation process. Instead of labeling all data, only samples that are of greater value for the learning process can be selected. As a method for the selection process active learning is implemented to alleviate the user. As a consequence the user is given an active role within the machine learning process, running on his device.

To what extent and in which ways can user privacy be provided and improved within the processing of mobile data on the user's device?

1.1.4 Wellbeing related to Mobile Contexts

A chance in mobile processing of human related signals is to see the diversity of emotional, social, behavioral and environmental contexts not as disturbing factor, but as source of possibilities. For that purpose, studies involving data collection, labeling, machine learning and demonstration are executed as an empirical objective in this work. The taken method is illustrated with the example of laugh recognition, drinking activity recognition and that of environmental wellbeing. Personal preferences in visual scenery are matter of interest in research of human computer interaction, as demonstrated in the work on route planning regarding visual preferences by Runge et al. [173]. In this work there has already been a relation between user

preference and environment, yet the objective measurement of the user's wellbeing in connection to environment has still to be developed. Where the surroundings of a city can have a negative impact on a person's wellbeing causing stress, other environments, such as forests, can have a calming effect. One of the objectives of this work is to make wellbeing objectively measurable in connection to the environment. The aim of the study was to evaluate the dependence of human physiology with different types of vegetation and buildings in the urban landscape. An interdisciplinary approach was adopted in the realization of the study, in which physical geography, human geography and computer science worked together.

Can mobile sensing be successfully used in interdisciplinary research on wellbeing in an environmental context? To what extent can personal wellbeing be measured in this respect?

1.2 Outline of the Thesis

This thesis is structured as follows, see also Figure 1.1:

- After introducing how this work is situated within the background of *Mobile* (Section 2.2) *Social Signal Processing* (Section 2.6) in **Chapter 2**, *"in the wild"* (Section 2.1) it is described, what aspects of *wellbeing* (Section 2.4.1) are used as common thread throughout this text. Related work can be found in **Chapter 3** that is followed by an overview on the structure and implementation of the MobileSSI framework in **Chapter 4**.

- In **Chapter 5** follows an evaluation of MobileSSI "in the wild" in the Affective Computing challenge of laughter-recognition. This reflects the exemplary transition of a core scenario from the lab into "the wild". Video is substituted by acceleration as modality to accompany audio. Moreover, asynchronous fusion is applied inter personal as well as multi-modal for a live demonstration.

- With **Chapter 6** the machine-learning capabilities of MobileSSI are extended to an interactive implementation of drink activity recognition. Data-labeling is realized with the support of a smart scale, evaluation is done comparing simulations of active learners to a fully trained model for drink-activity recognition. The chapter concludes with remarks on first user experiences of the interactive machine learning prototype.

- The influence of different environments and their respective local climate zone are examined using machine-learning on physiological data in **Chapter 7**. Self-assessed wellbeing is used as one basis of a machine learning model, while GPS-based segmentation of the traversed area is another source of labels for machine learning models. Results of a fusion of Skin Conductance and Heart Rate signals are presented.

- Concluding remarks can be found in **Chapter 8** summing up results of the individual chapters related to each other and the research objectives.

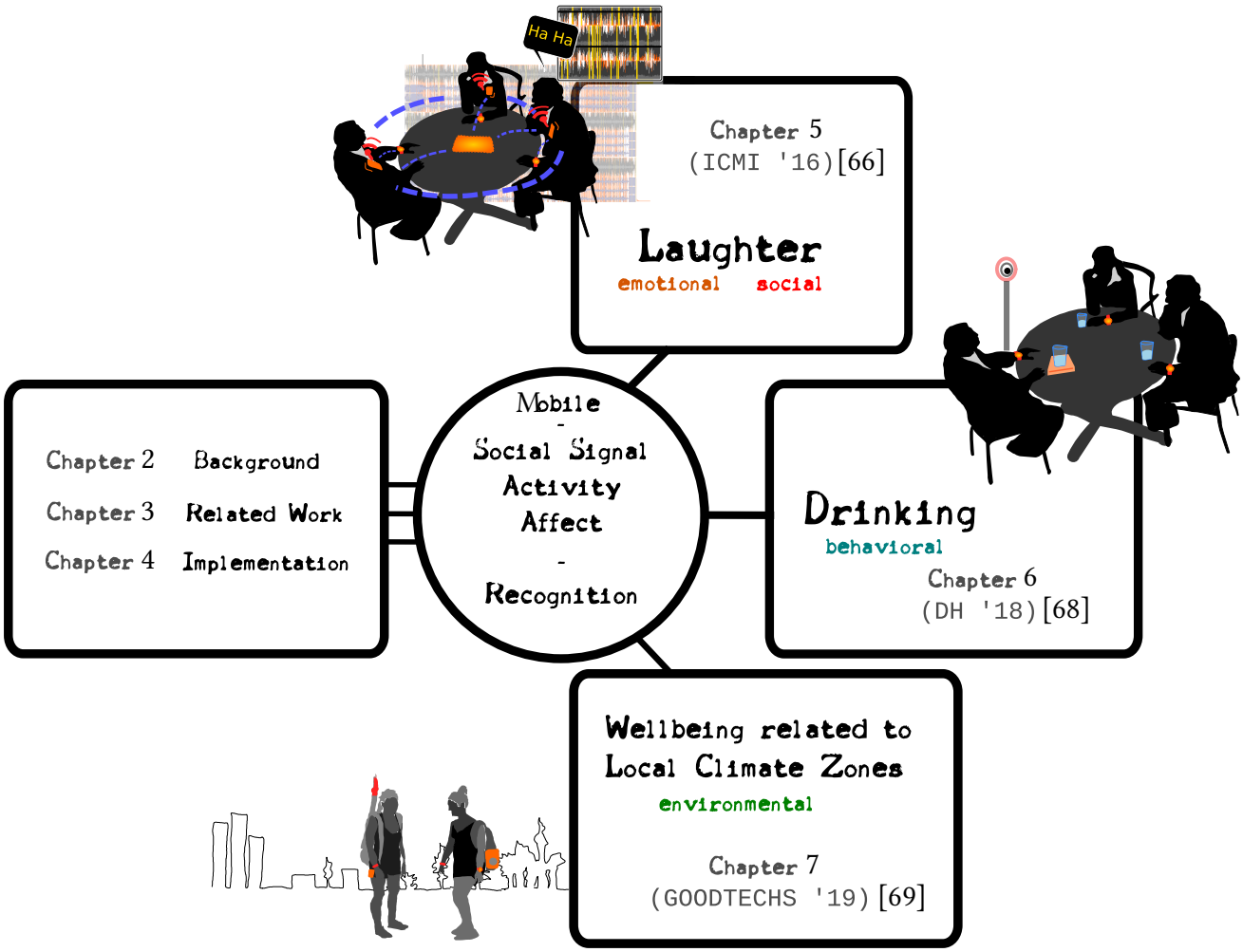


Figure 1.1: Outline of this thesis.

Publications

Some ideas, examples and figures included in this thesis have appeared previously in parts of the following peer-reviewed publications:

Publications directly related to this thesis :

Own contributions: software development, study design and execution, evaluation:

4th Chapter: S. Flutura, J. Wagner, F. Lingenfelser, A. Seiderer, and E. André "MobileSSI-A Multi-modal Framework for Social Signal Interpretation on Mobile Devices" *2016 12th International Conference on Intelligent Environments* **(IE)**, New York, NY, USA, 2016, pp. 210–213

4th Chapter: S. Flutura, J. Wagner, F. Lingenfelser, A. Seiderer, and E. André "MobileSSI: Asynchronous Fusion for Social Signal Interpretation in the Wild" *Proceedings of the 18th ACM International Conference on Multimodal Interaction* **(ICMI)**, New York, NY, USA, 2016, pp. 266–273

5th Chapter: S. Flutura, J. Wagner, F. Lingenfelser, A. Seiderer and E. André "Laughter Detection in the Wild: Demonstrating a Tool for Mobile Social Signal Processing and Visualization" *Proceedings of the 18th ACM International Conference on Multimodal Interaction* **(ICMI)**, New York, NY, USA, 2016, pp. 406–407

6th Chapter: S. Flutura, A. Seiderer, I. Aslan, C. Dang, R. Schwarz, D. Schiller and E. André "DrinkWatch: A Mobile Wellbeing Application Based on Interactive and Cooperative Machine Learning" *Proceedings of the 2018 International Conference on Digital Health* **(DH)**, New York, NY, USA, 2018, pp. 65–74 **– Best Student Paper Award**

7th Chapter: S. Flutura, A. Seiderer, I. Aslan, M. Dietz, D. Schiller, C. Beck, J. Rathmann, and E. André "Mobile Sensing for Wellbeing Estimation of Urban Green Using Physiological Signals" *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good* **(GOODTECHS)**, New York, NY, USA, 2019, pp. 249–254

7th Chapter: S. Flutura, A. Seiderer, I. Aslan, M. Dietz, R. Schlagowski, D. Schiller, C. Beck, J. Rathmann, and E. André " Interactive Machine Learning and Explainablility in Mobile Classification of Forest-Aesthetic" *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good* **(GOODTECHS)**, New York, NY, USA, 2020, pp. 6

Own contributions: software development, study design and execution, evaluation, where machine learning and physiological signals are concerned:

7th Chapter: J. Rathmann, C. Beck, S. Flutura, A. Seiderer, I. Aslan and E. André "Towards quantifying Forest Recreation: exploring outdoor thermal physiology and human well-being along exemplary pathways in a central European urban forest (Augsburg, SE-Germany)" *Urban Forestry & Urban Greening*, Elsevier, 2020, pp. 126622

Further selected Publications:

Own contributions: related work:

A. Seiderer, S. Flutura and E. André "Development of a Mobile Multi-device Nutrition Logger" *Proceedings of the 2Nd ACM SIGCHI International Workshop on Multisensory Approaches to Human-Food Interaction*, New York, NY, USA, 2017, pp. 5–12

Own contributions: software development, study design:

Y. Nakashima, J. Kim, S. Flutura, A. Seiderer, E. André "Stress recognition in daily work" *International Symposium on Pervasive Computing Paradigms for Mental Health*, 2015, pp. 23–33

Own contributions: ethnographic study, artwork:

I. Aslan, K. Weitz, R. Schlagowski, S. Flutura, S. Valesco, M. Pfeil and E. André "Creativity support and multimodal pen-based interaction" *2019 International Conference on Multimodal Interaction* 2019, pp. 135–144,

Chapter 2.
Background

This Chapter gives an overview on topics, that interact with the Mobile Social Signal Processing as a field of research and a method, as it is understood in this thesis. As such, it goes beyond my publications to date.

First there is research ”in the wild”, that examines phenomena, as they appear in real life, instead of recreating these phenomena in a controlled lab environment. Mobile Computing forms the technological foundation as well as the broader field of research, that encapsulates Social Signal Processing on mobile devices. It follows a model of wellbeing that connects to Affective Computing and Social Signal Processing.

2.1 In the Wild

The phrase ”in the wild” is seen as contrary to ”In the lab” and has its roots within cognitive psychology. Here observations in situ, made clear that models created in the lab did not hold up with processes, as they take place in real life [92]. Going beyond making observations ”in the wild” is Yvonne Roger’s approach [169] of developing ”in the wild”. Here designs, applications and solutions are elaborated with target audience within the target scenario. In the field of emotion recognition, ”in the wild” [206] is contrasted to ”in the lab” as well. Here the lab has advantages of higher quality setups and data, that often contain acted or alienated emotions. There are data sets of different kinds, that would be labeled as ”in the wild”. To fulfill the high amount of data to train deep artificial neural networks, data sets are generated from online video streaming services, containing self recorded data that are not acted. They might be described as data ”in the wild”. Viewed from the perspective of user interfaces (UI) computing systems nowadays are relying on many sensors, where video is only one, and not the most available one. A phone might still acquire information and run an application when carried in the pocket or while the user sleeps. Here acceleration and audio become the data of greater importance ”in the wild”. According data sets are are also recorded e.g. by the community doing

activity recognition or those ”big tech” companies, labeling private conversations recorded by their services to improve their speech recognition models. Since this thesis deals with work on a mobile machine learning back-end for wellbeing recognition, rather than full-blown applications, designing ”in the wild” is not a focus directly. Starting out with data sets recorded ”in the wild”, it extents to approaches generating models ”in the wild”.

2.2 Challenges and Chances in Mobile Computing

With the widespread adoption of smart devices a central aspect of ubiquitous computing as described by Weiser et al. [225] has become part of everyday living. Miniaturization of computers, better power supply with increased efficiency as well as advances in usability are important factors of the smart phones success.

When developing applications for everyday living, there are two factors that come to mind. Is it technically possible to run an application long enough, ideally a full day without charging? Is the software otherwise usable under the varying circumstances of every day living?

2.2.1 Energy-Efficiency

While Energy efficiency is necessary for long-term studies in science as well as deployed applications within every day living, advances are made in hardware (Moore’s Law), that enable us to run today’s complex computing tasks on a not so distant future’s generation of smart devices. A sufficient energy efficiency is a necessity to be able to use a mobile device in the daily routine. .

2.2.2 Loss of Control

Limited user control due to increasingly restricted input is an obvious matter of human computer interaction. At a lower level the software in mobile contexts is confronted with more complex circumstances, not foreseeable for the developer [52]. The user and also the developer is restricted in his device usage further, by the software-setup delivered in modern smart phones:

- environmental context influences input directly as in auditory noise, overly bright sunlight or time restrictions when on the go
- changing environments with different noise levels or characteristics of body motion while walking or using an a train

- application use might not be the user's be primary action when the phone is used for pedestrian navigation
- limitations due to operating system behavior (killing background processes for energy saving)

2.3 Persuasive Technologies

With accompanying us in everyday situations, mobile computing also holds the possibility to interact with our behavior unobtrusively [71]. In terms of M-health, immediate access to health-information without constraints of time or place is seen as a benefit, as can improved self-management of those suffering from chronic diseases [3]. Persuasive technologies such as gamification [122] and nudging [120] as well as reflective design [190] promise the creation of technology that invokes positive behavior change towards a healthier lifestyle, e.g. in supporting people to quit smoking [59].

M-Health chances an challenges can be structured as P5, as suggested by Gorini et al. [82]. This aspects of advancement over today's intrinsic properties of mobile technology promise a reduction in adoption issues.

- Predictive
- Personalized
- Preventive
- Participatory
- Psycho-Cognitive

Mobile Social Signal Processing, involving on-device machine learning can be used to create predictive, personalized and participatory apps. Predictive as in machine learning predicting patient's future health state and could then be used to prevent a disease by suggesting and monitoring behavioral changes. Involving the user into the creation of machine learning solutions leads to both user participation and personalization. Since Mobile Social Signal Processing handles input, and little influence on the app's content and behavior, its contribution to psycho-cognitive is limited by providing the extraction of the users' state of wellbeing.

A pivotal idea to this thesis is it to create technology that takes the stand of its user. There are a number of stakeholders involved in using mobile software with reference to health: data-driven companies, governments, health insurances, physicians and finally the user himself.

With the decrease of direct input and goals such as power saving, mobile devices can be viewed as agents as well.

Increasing the individuals' autarky [129] with respect to their data and the knowledge-transfer to his device, might turn out to reach goals in society that are impossible to reach otherwise. As a consequence of labeling data afterwards by foreigners, instead of asking the originators of the data for labels, can lead to misinterpretation or loss in information.

The circumstance, that machine-learning can happen locally in a feedback-loop with the user, that evaluates his model and decides what to share, might also increase the overall quality beyond what can be extracted from individuals participating passively in the process.

This leads to a conflict of interests within contemporary approaches to mobile machine-learning, where data are gathered on companies servers and knowledge is extracted from people via social networks and observed (online) behavior.

While capabilities exist, the interest of companies or states in terms of technology is wrongly seen as superior to the interest of the patronized individual, rather than being based on collaboration and common goals.

Viewing the owner, also represented through his device, as autarkic in the sense of:

- agency – taking part actively in the process of using an M-health solution
- collaboration – collaborating in the creation of M-health solutions involving the users' data
- data ownership – seeing the user as originator and rightful owner of his data

enables progress in health related mobile computing, to be a democratic dialogue of self responsible participants. Thus, a technological transition is proclaimed towards decentralized infrastructure that influences behavior also towards an active participation and technological awareness.

2.4 Wellbeing

Wellbeing has many facets, whereof only a few are topic within the scope of this thesis. While wellbeing might be defined as a state of good health and fulfillment, it again is a key feature in the World Health Organizations definition of health as a state of complete wellbeing [139]. The narrowest definition thus focuses on body and mind, from there a social aspect of wellbeing commonly is added, this is common ground to many wellbeing indices [121]. The addition

of social phenomena, is a step beyond the body and can be accompanied by a person's surrounding environment. To link individual components of wellbeing, and give an entry-point for interactive media, a behavioral aspect of wellbeing is added to this thesis model of wellbeing.

Thus, the wellbeing model structuring this work is divided in four parts. Starting with emotional wellbeing, that is body and self-oriented. Drawing on the potential of mobile devices to support us in everyday living the second building block is added focusing on behavior. Extending the view beyond one person concludes the wellbeing model in this work, by adding a social component and finally environmental wellbeing.

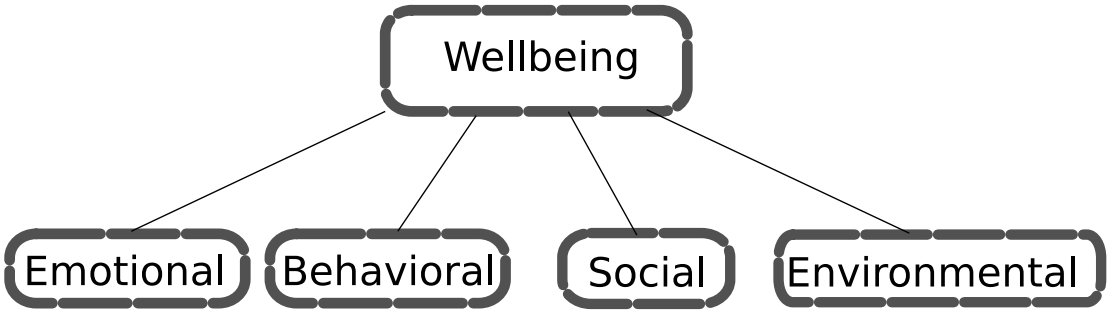


Figure 2.1: Wellbeing model underlying this thesis

"Action may not always bring happiness, but there is no happiness without action." – William James

2.4.1 Emotional Wellbeing

Deriving from a definition of emotions by Plutchik [154], gives an impression of the broad spectrum concepts of emotions might cover: *"An emotion is defined as an inferred complex sequence of reactions to a stimulus, and includes cognitive evaluations, subjective changes, automatic and neural arousal, impulses to action, and behavior designed to have an effect upon the stimulus that initiated the complex sequence."*

One can find the attempt of a feedback loop by the human acting emotionally and the stimulus. The scope of the emotion starts with neuronal arousal and subjective changes and extends to behaviors that have an intended effect on the stimulus. Where the behavior is what we witness, the inner state of arousal and subjective change is what we infer.

Thus, be it the right mix of emotions or the amount of positive emotions we show, emotions are central to the comprehension of a person's wellbeing [189]. The scope of the emotion starts with neuronal arousal and subjective changes and extends to behaviors that have an intended effect on the stimulus. Where the behavior is what we witness, the inner state of arousal and subjective change is what we infer [18].. While smiling and laughing are important behavioral markers for detecting a persons wellbeing, they also contribute to a person's wellbeing

by themselves. The concept of positive body expression, such smiling being the cause, not the consequence of positive emotions [94] has fueled applications in HCI [211] and thus form an important basis for related work and work presented in this thesis. Computers are a recent addition to the communication partners a person can have, which brings us to social aspects of wellbeing.

2.4.2 Social Wellbeing

In trying to define social wellbeing, loneliness could be seen as an antagonist of social wellbeing associated with family and community ties [88] within the concept of social capital.

This relates to mobile processing of social signals, since Computers are bringing people together successfully and contribute to our social behavior, not at last when looking at social networks. Then again the use of social network correlates with loneliness [30], also social behavior towards machines can be harmful [4]. This thesis' view on social behavior lies in another granularity though, based on its roots in social signal processing, natural human-human communication is a major inspiration. Instead of taking written, asynchronous communication or telecommunication as reference, the communication that goes with software prototypes of the work ahead are face-to-face, through audio or motion. This relates to research on how much we talk with whom in the "conversations monitor" [170] by Rossi et al. Reconnecting to emotional wellbeing, laughter not merely is an indicator of happiness but in contagious way highly social. On a footnote, laughter is flooding our body with hormones [21] that animate our appetite, which integrates well with the social aspect of eating and drinking that are also addressed within this thesis. While social wellbeing might be understood as a state, simple actions such as having a drink together can influence this social wellbeing, pronouncing the role of our behavior.

2.4.3 Behavioral Wellbeing

Following the thought by William James [94], that action is a requirement for wellbeing, behavior is the connecting point in this thesis' wellbeing model. One could argue, that drinking sufficient water, or non-alcoholic fluid equivalent, as well as exercising one's body should be categorized as *physical wellbeing*. Rather than distinguishing in mental and physical phenomena that can be seen as reaction to stimuli, contributing factors are identified. Healthy behavior and the change towards it is a chance of persuasive technology. This is underlined by B.J. Fogg [71] to whom behavior change is *"bespoken to have the potential of reshaping our live to be healthier and more fulfilled"*. Applications might be nudging [7] the user to cope with digital overload [137]. Nudging here is a tool enabling designs that influence the user subliminal towards a god decision. The "HappinessCounter" by Rekimoto et al. [211], is pervading everyday living with

the aim of changing our behavior in a positive way. It is a fridge that only opens to the smiling. Behavior also is of interest to the work ahead since it conveys action, that again is related to interaction with a computer system. This lets us sketch a future were the change towards a positive state in wellbeing and interaction with an ubiquitous computing system is one. The health application on our phone detects we are walking, carrying it is all it needs to interact with the system, the application estimates the benefit and might incite further action.

2.4.4 Environmental Wellbeing

Motivating people to action can be viewed as a central point, that makes an environment beneficial to a people's wellbeing. That is an perspective taken by using "Walkability" [74] as measurand, where grade to which an environment invites to walk through it, is captured. Limiting the set of environments to those, that are beneficial to our health, leads to therapeutic landscapes. Therapeutic landscapes are defined by Gesler as

"...those changing places, settings, situations, locations and milieus that encompass the physical, psychological and social environments associated with treatment or healing; they are reputed to have an enduring reputation for achieving physical, mental, and spiritual healing" [78]

Therapeutic landscapes, with the claim to healing can be found in spas, well-chosen places that are not necessarily nearby. While they are an important argument in favor of environments capability to improve our wellbeing, the places examined within the scope of this work are more ordinary and immediate accessible to a larger population.

Given the importance of multi-modality in this thesis, different channels in that an environment influence a person's wellbeing will be underlined.

That as much as the view from a window can have a significant impact on a patients health was found in research by Ulrich et al. [212]. This supports the work of Parsons et al. on scenic beauty [146], where is argued that, scenic aesthetic environments are holistic in an environmental psychological and cognitive scientific view, and therefore sustainable.

Beyond the visual scenery there are several modalities that on the one hand define a place and on the other hand make it possible to sense a place - be it using human or machine perception. An impression of how physical, cultural and embodied levels might be defined and sensed in soundscapes can be found in the following passage by George Revill: *"Sounds interact and mask each other high or low, loud or soft, incessant or fugitive. In spatial terms, heard sounds give embodied sensation to properties of depth, distance and proximity, suggesting feelings of clarity, delicacy and intimacy, transforming and animating the experience. Sounds envelope and reverberate deeply within bodies in ways which are specific both to their phenomenal properties and to historically constituted modes of listening, under- standing and interpretation."* [165]

Soundscapes are a field of interest in ubiquitous computing [171], where classes ranging from *forest* to *railway station* are recognized via machine learning as well as *brushing* teeth and *raining*.

While noise might not be the first association that comes to mind with pollution, the connection of pollution and air quality [31] might be more common. Modern cities have to fight smog, while spas have air quality as a key feature for recreation. Mobile computing can at first be used to sense air quality [87] and consequently might consider it as information while interacting with a user.

Another sensory-channel through which an environment affects us is heat [123]. While it is not a typical modality researched as a social signal, it is maybe the most direct link to climate. Climate within local zones, varies depending on buildings and vegetation found in that zone [19, 202].

When viewing environments as natural resources one might argue that they have to be guarded against consumption through humans. The "One Health" [47] combines the health of animals, humans and vegetation in an overall concept while maintaining sustainability. Future applications might consider a holistic approach to environmental wellbeing as well.

2.5 Affect Recognition

Affective Computing is referred to by Picard as computing that relates to, arises from, or influences emotion, also discriminated into the two classes of computers "being able recognize emotion, and to induce emotion" [148]. Affect Recognition focuses on the first of those two classes. Social Cues such as crying or laughing build a foundation of affect recognition, and the input side of Affective Computing. It touches on Social Signal Processing (SSP), described in the next section, in that it aims at achieving a natural handling of emotions in human computer interaction. As such it intersects with SSP, that has all aspects of natural conversation in focus. Next to input and output per se, an abstracted understanding of emotions is essential. This understanding is reflected in models of emotions, usually theoretical at first and implemented in programming logic finally.

Since emotions and the adjoined measurement of wellbeing are a cornerstone of this thesis, an introduction to models of emotions can be found in the following.

2.5.1 Models of Emotions

While it is conclusive, that affective computers have to sense emotions, understanding in empirical sciences goes hand in hand with measurements. Thus models of emotions were formed

and validated via measurements even as early as 1928. William Moulton Marston used blood pressure to find out if a person was mad or excited [119]. This resulted in his model of emotions described by Dominance (D), Inducement (I), Submission (S), and Compliance (C) and resulting into two dimensions: positivity and control. This contrasts in structure from simple models of emotion that start from discrete basic emotions such as acceptance, anger, anticipation, disgust, joy, fear, sadness and surprise [140, 154], but is already similar to the valence arousal model by Russell et al. [174].

While Marston reflects on submission versus compliance and the consequences for a society as a whole, Russell's perspective is more focused on the individuum. Nonetheless there is a continuity, e.g. dominance as axis in a multi-dimensional, continuous space of emotions [175].

Alignment of discrete emotions in that space might vary from culture to culture and also on the emotion and its term. Focusing on basic emotions [57] promises robustness e.g. across cultures. From a practical standpoint continuous spaces are sometimes divided into discrete classes for valence it might be: "positve" "neutral" "negativ" to cover a broader emotional space than discrete emotions such as "happy" and "angry". An overview on computational models of emotions can be found in the work of Marsella et al. [118].

Annotation in a three dimensional space of valence arousal and dominance is a demanding task that requires a tool. A contemporary approach to emotion-annotation unifying discrete emotions, multiple dimensions and supplementary annotation in free-form, is the Geneva Emotion Wheel [176] in Figure 2.2. Noteworthy here is the stance of the observer. Am I observing my-

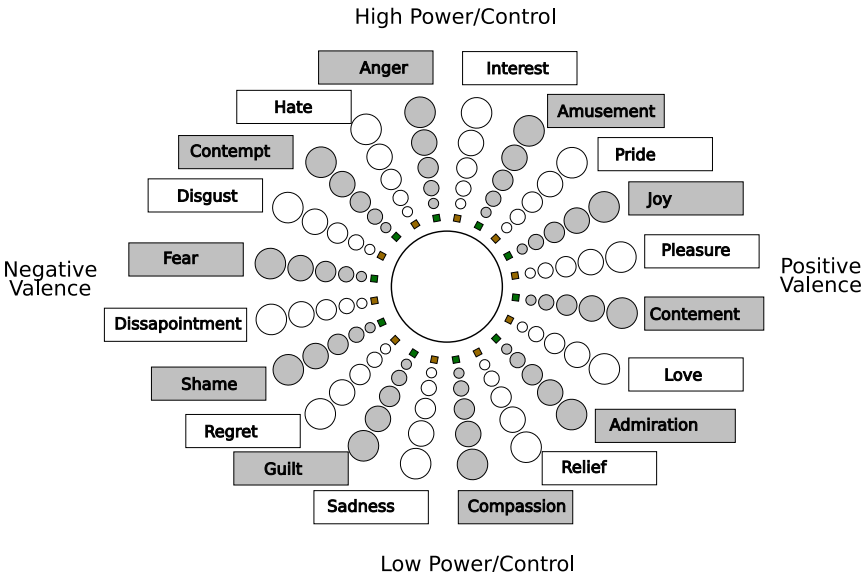


Figure 2.2: Geneva Emotion Wheel 3.0 [176] ("no emotion felt" and "other emotion felt" are additional options of the Geneva Emotion Wheel missing from this Figure)

self or is another person doing the job? Is the aim to design a system that can recognize what an by-stander would be able to see, or is self-perception the target, so where lies the ground truth?

The capturing of emotional states is an important ingredient to SSP and Affective Computing. This can happen by:

1. self report e.g. using the Geneva Emotion Wheel in Figure 2.2,
2. labeling by one or multiple experts or
3. retrieving the label information directly from the stimulus that induced the situation that was recorded.

With a shift to "the wild" it is increasingly hard to have recorded data, that speak enough for themselves to make option **2.** viable. Option **3.** is also better suited for lab settings where environment and stimuli can be controlled. Self-report (**1.**) therefore is the obvious choice for mobile labeling processes, not restricted to emotions only. In a mobile setting it might be desirable to simplify the annotation process and therefore reduce the complexity of the underlying model. Here valence [14] is a favorable choice for an overall impression of the user's wellbeing, since it is a component that can be found in most models of emotion. Thus it is employed later on in Chapter 7.

2.6 Social Signal Processing

The community of Social Signal Processing (SSP) [147] is driven by the idea of natural human communication with computers. Thus, one focus lies in analyzing human communication, identifying how messages are transmitted. The obvious here is the spoken word procesed by speech recognition that is a field of research by itself. Speech, at closer look, is accompanied in natural communication by additional channels, referred to as non verbal communication. Non-verbal communication can by itself transmit meaning, underline or change the meaning of verbal communication. For a computer to become a more natural communication partner it has to sense actively rather than passively waiting for instructions. The active sensing is realized via digital signal processing and in most contemporary cases uses machine learning for information extraction.

2.6.1 Social Cues

To be able to process conversations automatically, the stream is broken down into a series of events. Those events are named behavioral or social cues, since they convey a part of the information that can be extracted by combining all cues [58, 217].

As such, a social cue is not an entity with clear boundaries and content, but rather a beat that overlaps with others, a modifier for other information channels or a hierarchical part of a larger construct, e.g. a joke.

Those cues can be categorized into five different types of signals they are sending:

- State of emotion: cognitive, attitude.
- Manipulators: towards the environment or oneself.
- Cultural emblems: common only in a certain cultural circle, such as ”high five”.
- Illustrators: underlining information transmitted in other channels of communication.
- Regulators: affirm other communication partners or indicate turn-taking.

Verbal communication is naturally combined with non verbal cues that can fulfill different roles in changing meaning or organizing flow of communication. For an orientation towards wellbeing the emotional and cognitive state of a user can be considered the most valuable information.

2.6.2 Role of Context

A further aspect of natural conversation is, that it is not necessarily self-contained. Next to aspects of nonverbal communication that manipulates or supplements explicitly spoken communication, there is a wide range of context a conversation might refer to. This is clear within a task-oriented dialogue [36], where the task defines requirements ”slots” that have to be filled and the user’s intents can be determined. In the wild contexts are richer and more fluid. Social interaction can vary depending on environment and main activity, e.g. walking or drinking, but also depending on the emotional state.

2.6.3 Modalities

The findings of McGurk [124] led to relying on multiple modalities when analyzing human communication. In case of ”hearing lips and seeing voices”, visual cues are examined in combination to voice. By combining lip-movement not fitting the original ”ba” but ”ga” our auditive perception seems to adapt to ”hear” ”da”, a mix from both senses. Consequently, the interleaving of different senses is an important part of our perception. Since SSP tries to engage computers in natural conversation, it should not rely on one sense also. The information processed by a single channel of human communication typically is called a modality.

Processing a multitude of modalities, rather than a single one, also promises advantages in reliability. In the extreme one information channel is missing completely, e.g. visual input due to bad lighting condition, that does not lead to information loss since the other modality, e.g. audio, transmits information still. The modalities might cooperate by redundancy or complementary as illustrated by André et al. [8]. To improve recognition, multiple modalities can be

combined in a fusion process. Wagner et al. [220] moreover discuss the more complicated relationship of multiple modalities, on natural data, due to contradicting social cues. Similarly the cooperation of several people might support or disrupt each other. Thus, in scenarios involving a group rather than a single person, the relation of individuals' behavior e.g. synchrony [214] can convey social cues.

2.6.4 Corpora

To develop and test SSP algorithms, data collections are required. Those data collections are called *corpus*. A single recording from a sensor consists of samples, since analogue signals have to be digitalized into a sequence of discrete values. Sensors can be sporadic by sending an event for every sample or using a fixed sample-rate, generating samples regularly at every even time-interval. Different sensors' data might be recorded in different *tracks*. All recordings that happened in one go form a session. The people involved in the recordings are usually called *user* but they might take different *roles*.

The recorded data have to be *annotated* with experts' knowledge based on a *ground-truth*, that might be extracted by viewing audio-visual recordings or by observations. Those annotations can be done in different *schemes* e.g. as continuous numbers on different axis (e.g. for regression learning), per defined *labels* for classification or as free annotation. The process of labeling might be executed by several annotators and lead to a gold standard.

Corpora are indispensable for research, especially as long as it has an explorational component searching for new solutions. This has to be distinguished from final products and later stages where sharing aggregated knowledge in models rather than full data collections is a viable option [234]. Corpora could be basis for scientific methods other than Social Signal Processing, e.g, manual transcription, the further steps along the process, namely training distinguish SSP further from other approaches.

2.6.5 Recognition Pipelines & Training

Roughly, a machine learning system is composed of two phases, training and recognition. A recognition *pipeline* that is running in real-time is important to be able to interact with the user. *Pipelines* name graphs of data processing flow, where different stages of the process can happen in parallel. Thus new sensor data can be recorded while old sensor data is still processed, e.g. in a machine learning model for classification.

The machine learning workflow is necessary, to create the model used within that recognition pipeline.

- Sensor

> Filter & Features

> Classifier

At the beginning of a recognition pipeline, a sensor converts a continuous, analogue signal into a discrete digital signal. This is processed using filters and feature extraction to be fed into a classifier that extracts information from the data stream.

The classifier contains a model that resulted from a machine learning workflow. First in that workflow is the recording, usually done with the same sensor that would be used in the recognition pipeline.

> Record

> Label

> Learn

The data have to be labeled with the information, the classifier should extract later on. Usually the labeling is based on a ground truth from which annotations are created by one or more experts. In a training process the machine learning model is created, that later will be attached to the classifier. Part of the training is also testing or evaluating to judge the quality of the model before deployment.

With multiple sensors and modalities that are fused at different stages of the pipeline and the combination of recognition and learning processes, the workflow is reshaped later on in this thesis. Also training, evaluation and recognition are interleaved.

For implementation details refer to Section 4, an outline can be found at the end of Section 3.

2.6.6 Real-Time Recognition

For user interaction systems have to be designed to rely on the current data instead of a whole data-collection. This means, that processing takes place on a stream of data, without the ability to look far into the future or to access all old data, but working iteratively.

Typically this is realized in the recognition process by using a sliding window approach. Data are cut into chunks of the moving part *"frame"* and an overlapping part *"delta"* or *"context"*. Thus, a result is presented every frame and creates a responsive system to interact with. This does not mean that hard time constraints are given and therefore processes consuming more compute time will not be interrupted. A slow system thus might get increasingly inresponsive, as the work load piles up.

Offline processing enables a detailed analysis of data, without the need for high performance and restrictions regarding random access on the processed corpus and is used in several software solutions such as Praat [27].

2.6.7 Incremental Learning

While real-time recognition is a common approach in SSP, the training process usually happens offline. In incremental-learning, training does happen in smaller batches, without the possibility to iterate over the whole data collection. Offline, models are often trained in one go, without warm-starting on an already existent model nor is the process interactive, incremental learning does not need to preserve all data that result into a model and enable training while observing, in cases where recording of a ground truth is hard. While data collections could be created with annotations from observations, incremental learning has the advantage to also enable iterative evaluation in place. Thus, systems based on machine learning can be more agile.

2.6.8 Social Signal Processing as Methodological Approach

In summary, SSP provides a framework to extract information from natural social behavior via sensors. This goes beyond mere factual structure but also involves affect. Human behavior is induced via study design, e.g. causing stress via a Stroop-Test [164], recorded and annotated. As such SSP usually provides a wider believable context to the particular behavior that forms the research objective, and strives for real world applicability, also by providing real-time feedback. Those methodological characteristics are common in Chapters 5, 6 and 7, while the focus is extended from social cues to behavior and wellbeing touching on Affect and Activity Recognition.

Chapter 3.
Related Work

Figure 3.1: Mobile Lovers (Modern Communication) based loosely on work by Banksy.

Mobile Social Signal Processing (MSSP) at first is the continuation of Social Signal Processing on mobile hardware. Hardware evolution and miniaturization enables Pocket-PCs to handle the workload that once needed desktop or main-frame machines. For the study of using natural communication towards computers, mobile phones seem a obvious choice, since they accom-

pany us ”in the wild”, where people interact naturally. Unobtrusively of devices, sensors and the sensing-setup has to be considered when examining natural behavior ”in the wild”. This is important e.g to not restrict natural motion by wires or weight.

This Chapter embeds my contributions in a wider context of other approaches to communication channels and goes beyond my publications to date.

There are two sides to combining SSP and Mobile Computing. Multimodal interfaces [142] are a better fit to the requirements of changing contexts, than traditional input. Also, natural communication has become more digital by itself [216]. We might even go so far as text each other even if we are in the same room, or make a phone call instead of searching for someone analogously at a crowded place. Consequently, digital communication is studied as social signal by a variety of approaches to Social Signal Processing on mobile devices that are e.g. examined by research of Palaghias et al. [143] and outlined by Vinciarelli et al. [218].

Mobile Affective Computing [155] addresses challenges ranging from detection of emotional states to personality trades.

Mobile sensing technology is considered to have a key role in studying social factors in behavor change towards a healthier life-style [2]. Aharony et al presented a year-long study in a lving-laboratory monitoring physical activity as well as digital media use, while presenting subjects with incentives to engage in more pyhsical activity.

3.1 Communication Channels in Mobile Social Signal Processing

While the work at hand considers only information from sensors (see 3.1) for the extraction of social clues, the broader picture of Mobile Social Signal Processing, it is embedded in, is presented in the following. The taken approach thus can be motivated and defined. In the following an overview on communication channels in MSSP as well as further challenges within environmental contexts, affect recognition and current research topics with shared goals such as crowd sensing. This chapter is concluded by presenting frameworks developed for MSSP and their features in relation to MobileSSI.

Meta Data

Starting from a high level of abstraction, there are approaches in Mobile Social Signal Processing integrating meta data as data source. Those are attributes raw data, instead of direct streams of information. Meta data might e.g. be how often calls were made and of what duration. Which app was used over what amount of time [181]. Also contacts and chat protocols can be source of meta data, where not the actual content is subject to examination, but rather statistic figures:

how many contacts, how is the variance in communication frequency within those contacts and how dense are the contacts interlinked. Those data are readily available in digital form and can therefore be parsed easily by applications as done by Burke et al., who modeled aspects of social capital depending on social network parameters such as friend count [30]. Meta data can also be found in application usage and call logs that are considered e.g. in boredom detection by Pielot et al. [150]. Meta data typically are restricted to digital communication and due to its slow evolvement has not to be processed in real-time. Thus, it has not been subject to the studies and implementation of this work

Text

While text messages also are already digital, in contrast to spoken language for example, natural language processing, that is needed to gather digital meaning from natural conversations, is a difficult problem to solve. Personal communication, "in the wild" is increasingly handled via instant message services, that can be linked to locations. Consequently it is relevant to mobile processing of social interaction. It is common for MSSP-Frameworks to use text messages for information retrieval [130]. Furthermore, there is a variety of work on sentiment analysis, e.g. using emoticons as labels to classify the twitter posts based on word sequences or *n-grams* as realized by Go et al. [81]. The sentiment analysis of text messages extracted from social networks also serves to investigate the influence of urban greenery on the emotions of visitors. Studies based on data from New York and San Francisco show, that visitors of urban green are not only happier, compared to people in other places of town, but also stay happier for several hours [153, 184].

Network Activity

In a more indirect manner, mobile devices can be used to identify people (if WiFi or Bluetooth is enabled) and calculate their position [97] in relation to other objects [99] or within places. While network activity can be viewed as an abstract representation of human behavior, in a similar manner to meta data or written text, it might also seen as a sensor to measure natural human behavior, e.g. the proximity [33].

Social Signals not Involving Processing of Digital (Social) Media

As mentioned beforehand, while introducing Social Signal Processing, there is a variety of "social cues" such as laughter, that are consequently basic concepts in Mobile Social Signal Processing as well.

Those cues are transmitted in different communication channels. Underlining of a message can be realized via a hand gesture and thus body motion or paralinguistic in speaking up. Both can be captured via mobile devices. In the following an overview is given on body and face related social signals, paralinguistics and physiology. With physical activity a different concept is presented that is less focused on social interaction but delivers tools for processing of the same data on mobile devices.

Body Related Social Signals

Body related signals are presented in the following from the least fine granular view of proxemics to postures and more fine granular gestures. Proxemics form a high abstraction, since persons are looked at not as a detailed state of body, by their mere position. Distances between people underlie continuous negotiation (how close am I to you?) and thus are a relevant social signal. The space between people depends not only on their relation towards each other but the social frame and place they inhabit also, imagine a crowded market compared to a few people waiting for the bus. The proximity is an important measure also for face to face communication and considered a key variable in the spreading of behavior change [30]. Proxemity might be sensed via dedicated devices [34] or using signal strength (RSSI) of wireless networks [33] such as Bluetooth Low Energy. This has become relevant recently e.g. within Corona warn apps, since proximity influences the spread of the COVID-19 virus – thus the recommendation of "social distance".

Back to social signals as a whole, not only the position per se is of importance, but also how our bodies are positioned in postures. Are we leaned to or from each other, turned sideways or facing each other.

Posture as social expression might contain crossed arms as closed, and dangling arms as open position and thus adjoin with gestures. To a certain degree this information can be extracted using position sensors in mobile devices [90].

Posture next to inter personal communication also is relevant to our state of wellbeing. Sonification of the posture of a persons back can raise awareness the person's physical capabilities within a rehabilitation process [199]. In this case posture is not used explicitly as a social cue within a dialogue and as such is not immediate matter of interest in SSP. Yet postures are a part of noverbal social behavior and acknowledged as social signal.

While postures are more static, gestures typically capture movement. Gestures are important ingredients to lively conversation and are connected to personality trades such as extroversion [72] or culture[60]. In this work gestures have to be distinguished from motion-symbols that trigger features in applications such as mouse-gestures [172]. On smart devices [20] the gestures in MSSP tend to be between the extremes of application defined and natural interpersonal

behavior. Gestures follow a similar sequence in every execution and often are matched with pre-recorded gestures using approaches such as FastDTW [178]. Furthermore a single gesture is limited to a part of the body in contrast to postures, and therefore a motion tracking device monitoring a single point can be used for gesture recognition, e.g. an accelerometer. Gestures are a common concept that is easy to grasp and implement when abstracted to drawing a circle, yet complex in the natural interaction with verbal communication and consequently still subject to research in Human Computer Interaction [177].

Face Related Social Signals

Compared to stationary machines that face us most of the time during interaction and thus are able to recognize our mimic, the role of mimic changes when using mobile devices. Since mobile devices accompany us on the body most of the time, they might capture more of our everyday life, but might capture less of our countenance, than traditional personal computers. Even though mobile phones are equipped with extra sensory namely front cameras, they have to rely on other channels of communication, e.g. motion, most of the time, since they accompany us in our pocket or hand bag. Face related signals yet most commonly are extracted from Image data. Even though video processing is a resource intensive task, recent development has led to real-time recognition of facial landmarks [10] on mobile devices [114].

A further face related social signal is Gaze: Is the dialogue partner interested in the current topic, is he lost in thought? Whom do we intend to take the next turn within conversation. Similar to mimic, the typical use of smart devices in analogue communication scenarios, limits smart devices ability to capture gaze. Nonetheless there is a trend in MSSP to detect gaze without using hardware especially build for that purpose [41]. To employ gaze detection on the go, smart goggles are developed and considered [209], that enable gaze recognition without bulky devices, that limit users' movement.

Paralinguistics

Speech recognition might be seen as pre-processing for text and sentiment analysis in natural language processing, since it generates text from audio. Social Signal Processing often focuses on information that is lost when just considering a transcript from speech to text. For example, intonation, prosody, gaps and fillings, along with laughter, can provide insights into mood, personality [43, 91, 117] and a variety of social aspects in speech. Those aspects of behavior distinguish verbal social cues from neutral speech, as used by a reporter reading the news on TV.

Even though the point of view changes from sitting opposite of each other, to being worn on the body for most of the time, in the use of phones compared to the desktop computers, speech

stays a reliable communication channel and therefore a modality of central interest in Mobile Social Signal Processing.

Physiology

Physiology is rarely considered a natural channel of communication, although we can certainly see when someone sweats under stress, or hear when a person's heartbeat or breathing increases in intimate situations. This might be due to the fact that we rarely use physiological channels consciously. Nonetheless, physiology can be used to deduct a persons state, that is relevant within a social context, and therefore is considered part of SSP. Wearables (smart watches, fitness-bands) have skin contact and thus, physiological signals are becoming a common input in commercial-grade devices e.g. in relation to sports training and physical activity. Using physiology for emotion recognition has been realized early on by Picard et al. [149] based on muscle activity (EMG), pulse (BVP), skin conductance (SC) and respiration (RSP). Physiological data are hard to interpret, compared to video since they do not stand for themselves but need context and medical knowledge to interpret. Morover, physiological signals have a tendency to motion artifacts, if they are form an unobtrusive source without glued electrodes. This is especially the case in commercial grade fitness wrist bands, that are relevant for measuring wellbeing on the go. Thus, BVP commercial grade physiological signal acquisition has been integrated into MSSP-frameworks for unobstrusive sensing [103], while bulkier GSR and ECG, requiring adhesive electrode [12, 61] are used due to the higher data quality. Communication in both cases is typically realized via Bluetooth. Compared to audio the data acquisition is a restriction only to the wearer's privacy and not the conversation partner's privacy as well.

Physical Activity

As already indicated, social cues, such as postures or gestures, can be seen in a broader context as activities of daily living. Activities such as drinking and walking or sitting might accompany conversation but can occur disconnected as well without being a social cue. Activities in the sense of activity recognition [162] can be seen as a super-set of both postures and gestures, combined with all not necessarily social activities and thus do not integrate into the concept of communication channels and social cues. Brushing your teeth for instance can be recognized well using a wrist worn accelerometer but is rarely a social activity. Activity Recognition has been integrated into MSSP-frameworks to provide applications with context such as, if the user is walking [12, 12, 103, 112, 130, 224].

Where as physiological signals are influenced mainly by physical activity and therefore can be used in activity recognition [233], they are associated and used to detect mental load also in recognition of emotions, mood or stress [204].

3.2 Emotions, Mood and Stress

Rather than being communication channels, emotions mood and stress are concepts on a higher level of abstraction that can be traced via different modalities and communication channels. This makes them a topic in MSSP nonetheless, since emotions are expressed via prosody in paralinguistics and here are highly relevant information in conversations. Emotions recognition, as mentioned in Section 2.4.1, also has a key-role when it comes to wellbeing focused applications, be it by inducing positive emotion and recognizing them within a feedback-loop or by affect recognition for the creation of emphatic machines or mood diaries.

Mood and Stress usually spread over a longer period of time, compared to emotions, which make them more attractive to long-term studies. While emotions might be induced in lab conditions, the observation of mood and stress "in the wild" is more adequate [11]. Long-term assessment is also what links MSSP to M-Health. Here applications acquire the ability to provide features for psycho-cognitive involvement in a person's "lived experience of illness" [82].

In the case of mood and stress recognition pursuing a third person view is not preferable, instead of relying on a ground-truth such as video and the labels of an expert, mobile solutions have to increasingly rely on labels by the users themselves, since no ground-truth and therefore no objective label of an out-stander exists. This shift of perspective comes hand in hand with the shift of sensors used.

Here has to be distinguished between a sensor and the modality perceived, since a body's motion can be recorded using a camera from an bystander's view or via accelerometers attached to the body itself. For example, speech might be recognized based on brain-activity instead of audio data [132]. Physiological signals can be used to recognize activity as well as stress [61]. Communication channels as seen from a theoretical, human centered background do not always match the sensors used in a (M)SSP-pipeline.

3.3 Role of Context in the Wild

As mentioned earlier, context plays an important role in natural social conversation. We naturally refer to our environmental "put that there" [28] but also refer to situational context. Social interaction might be executed on the go in a wider sense of a shared activity e.g. cycling or hiking. To interpret the shared activity of cycling as a dialogue is a rather far stretch for natural understanding of a dialogue. Nonetheless, it is social interactionsun2010activity. Imagine a agent accompanying you on a cycle trip, making remarks about points of interest or asking for your wellbeing while approaching a steep section of the route. Would you call that agent social? Mobile Social Signal Processing focusing on input is bound to tackle corner cases that would not be considered matter of social behavior in a lab setting.

3.4 Environmental Context

While environments might be used in natural conversations as a reference, they also might influence conversations by noise or similar disturbances. Those noises can overlay recordings and thus are a challenge for filtering. However they can contribute context information as well e.g. via noise mapping [159] in urban environments. Here the characteristics of different noises are mapped to e.g. urban areas. Auditeur [136] and SoundSense [224], as MSSP-frameworks provides the envronmental noise level as context information, where SoundSense [112] uses audio to distinguish indoor and outdoor environments within detecting walking-activity. In the wild, there is little control over the setting, thus the software has to cope with the challenges arisen and try to benefit from circumstances, where possible. Beyond references in communication, environments contribute by large to our wellbeing. The environment influences us - activates, stresses or relaxes us. This information can be extracted with the help of sensors (which measure our physical reactions). Above all, this could be a valuable insight for preventive M-health technology.

3.5 Further Framework Features

Aforementioned communication channels and contexts are core features of MSSP-Frameworks. In the following further features, that work under the hood, are described. Also aspects like multimodal processing, privacy management, machine-learning approaches and rapid prototyping can be seen as technical details, in the case of interactive machine learning and privacy aspects, they can nonetheless considerably influence the role and interaction concept of the software. Thus, they are described in the paragraphs below.

Multimodality

Multi-modality, the synchronized processing of multiple data streams from different sensors, is central already to SSP in the lab. It has arguably become even more important in mobile settings to expand a system's accessibility for non-specialist users and enhance the robustness of the system [142].

While multimodal interfaces see multiple modalities often as choices how an input can be generated in different situations, needs or preferences, in (M)SSP those modalities are processed at once, when possible. Multimodal approaches can be introduced at different levels of the framework, but mostly affects different stages of the processing pipeline at the level of calculating joined features or fusing different models.

The variety of sensors has to be integrated. Their recorded signals have to be handled in relation to each other, e.g. by syncing the data streams. With different signals, according features and fusion approaches have to be engineered. Sensors can be used analogue to senses, such as microphones for paralinguistic utterances, or across in the case of physiological signals for activity recognition. This thesis employs fusion of signals that are analogue to senses with signals that go beyond their analogous use. Sensors are selected to be useful in secondary use of smart devices, e.g while walking, and thus are as unobtrusive as possible. Conclusively, multi-modal support stretches across the frameworks design and is not a simple addition.

Privacy Approaches

While data-ownership is a formality in study design and execution in the lab, MSSP aims for approaches that could be employed in the wild, where ubiquitous recordings are bound to conflict with privacy. Furthermore, models that are not created to explicitly respect privacy and consider fairness tend to discriminate against people based on unethical features such as gender or skin color [54, 55, 96]. Unfair models can be result of data-sets that are created without explicitly balancing data according to person related features, or by not dismissing features that are considered unethical.

To counterbalance the problem of software disregarding privacy, different approaches can be identified in mobile frameworks. On the one hand processing can take place on mobile-devices solely [112] on the other hand information can be filtered and encrypted to contain only as much information as is needed and can be handled securely within the distributed processing pipeline. Both approaches add challenges to the creation of a MSSP-framework.

Considering privacy at a feature level introduces additional filtering [238], that has to be checked for integrity. Advanced algorithmic reconstruction of reduced information might be possible, as a counter measure to minimizing features. Moreover, as a result models quality might suffer from the removal of information in the feature vector.

Local data processing has to cope with the hardware and software limitations of smart devices and yet might be compromised by other applications such as operating systems or malware compromising private data. The framework Auditeur, e.g. enables on device processing for private spaces and requires user authentication to access private ”soundlets” [136].

Those privacy approaches are limited mostly to the prediction process, without creating new machine learning models. An exception is the use of unsupervised learning in mobisys [112], where no human labeling effort is needed.

Interactive Machine Learning & Labeling In The Wild

A consequence of the perspective shift from labeling by an observer to self-report, is a change in the machine learning approach. Instead of involving experts that deploy and solely supervise the machine learning process, the end user is integrated more tightly in the process [5]. This starts with data-labeling [213] and leads towards model validation and pipeline creation. The work at hand focuses on interactive model creation and therefore labeling. Interactive machine learning therefore involves machine initiative (Active Learning) and responsiveness, in a perceivable change in the behavior of a model.

When involving users into labeling there is a variety of approaches imaginable:

1. stimuli induced labeling – as used in the lab, experts are involved in designing the setup that induces behavior of a certain label in the user,
2. user initiated labeling,
3. labeling from behavior and interaction, with smart objects
4. Active Learning – the model requests labels by theire algorithmic value,
5. mixed initiative cooperative learning.

While **1.** - **3.** work within a classic machine learning setup, **4.** & **5.** involve also the initiative of the system in requesting labels. This is desired in order to reduce the labeling effort, limiting it to data that are most relevant to the learner. To integrate labeling into natural behavior the use of smart objects can be beneficial (**3.**). In Chapter 6 a smart scale is used as smart coaster to provide labels for drink activity.

In Active Learning it is assumed, that labeling is a pricey effort accomplished by asking an *oracle* [192]. The samples are therefore selected for the maximal information gain to reduce the cost. Within MSSP the user has to play the role of an oracle in that sense.

For actual user interaction several additional constraints that have to be taken into account. The user has to be provided with feedback, that enables him judge the quality of the model. Therefore, it is a necessity that the employed algorithms are responsive too. The feedback must also be provided in a manner that suits the user, queries have to be filtered to not be annoying without diminishing the information regarding the underlying system behavior. Interactive machine learning on mobile devices consists of several challenges, where some are algorithmic and some lie in the design of user interfaces [182]. With the reduction of graphical user interfaces in mobile devices such as smart watches, movement based interaction gains in popularity, interactive machine learning is here seen as a key technology to design interaction [80].

Rapid Prototyping

Closely connected to user interfaces for data-labeling on the go and custom pipelines is rapid prototyping. Rapid prototyping describes a method to quickly develop drafts, in the case of computer science, of applications. Typically, this is especially true for user interfaces, that can be created interdependently from application logic in different versions and iterations [73].

MSSP-recognition pipelines form user-interfaces themselves, and thus there are mechanisms that allow their prototyping. Rapid prototyping can be relized through pipeline configuration via flow diagrams [114] or config files.

Since the user is more tightly involved in the creation process of MSSP creation of traditional mobile UI [111] is an important part of the prototyping process as well. Web technologies and corresponding communication-protocols such as websockets can here be of use.

The resulting custom interfaces help to provide the user with visualizations and feedback for individual applications and MSSP approaches, that are fit for different scenarios and makes interaction with the application smoother. This also includes a custom way of user input that is adjusted to different machine learning problems, e.g. the labeling process.

Subsequent Topics

Beneath MSSP there are further research topics that share central aspects and solve similar problems in hardware, algorithms and field studies. Viewed from the analysis of conversations via text messages as mass-communication the human as a sensor [42] can be seen as a technologically related topic. The term crowd sensing [83] [84] [32] follows a similar thought, where people are not directly used to sense for events, but information mined from mass-communication as well. While text messages within social networks are typical data to be processed, other data, e.g. location via GPS, images shared online etc. are also common scenarios.

When it comes to intelligent environments and being instructed to use different sensor sets depending on location and situation, the topic of MSSP is joined by opportunistic computing [37]. Complex Event Processing [229] copes with similar processing pipelines as (M)SSP and as such can be referred to for data processing and machine learning solutions. In addition, continuous signal processing is used in M-health and E-health applications [138], which share emotion, wellbeing and environment as topics in conncetion with MSSP and are partly based on the same software frameworks.

3.6 Conclusive Overview

SSP has the goal to add natural communication cues, such as affect and emotions to human-machine-interaction. With the ability to conduct long-term experiments and user interaction with mobile devices, this expands experiments from tackling short-term emotional states, towards mood and stress e.g. by Ertin et al. [61]. Together with the ability to recognize environments and their influence on the user, mobile sensing based on MSSP is a tool that can provide solutions to challenges in the domain of M-Health [82]. Using mobile phones, it is close at hand to analyze phone conversations, as done in the dissertation by Anna Polychroniou [156]. Today a lot of private communication already involves digital media that is increasingly analyzed in MSSP as well. The approach of this thesis is to go a step further and to study natural behavior not directly involving a smart device, but rather seeing it as an unobtrusive companion on the body.

The direct perception of social cues is supplemented by the recognition of higher-level concepts underlying communication. Higher-level concepts can for example be emotions, which can be identified using various approaches, e.g. by using physiological signals. The context of interaction, activities and environments plays a bigger role in research with mobile devices, what leads to the exploration of corner cases between different research topics, such as Activity Recognition that can be applied to drink activity as well as on laughter.

Sensing ”in the wild” inevitably takes place in spaces that are more private than an artificial laboratory environment and at the same time relate to something more personal than a stationary computer does, which is limited to the location of the user’s office. Privacy issues such as ”discrimination of individuals based on private personal features”, due to imbalanced training data or wrongly chosen features [55, 96] are to be avoided by design and user control.

Integration of interactive machine learning (iML in Table 3.1), is rare in frameworks of mobile sensing, which lets us inferre, that infrastructure for the local and therefore, privacy-compatible, creation of machine-learning models is missing. MobileSSI is providing customly designed user interfaces for labeling on the one hand and an interface for active learning on the other hand.

Topics such as crowd sensing, that focus on a larger data-pool but share a lot of principles and challenges in data-processing and according to Capponi et al. rely on similar frameworks [32] as Mobile Social Signal Processing, even though the perspective on the matter changes.

When reviewing frameworks in contemporary surveys from mobile crowd sensing [32], continuous signal processing for M-Health [138] and MSSP [143] it shows that a significant part of the frameworks in literature target legacy systems (Symbian). The frameworks are rarely cross-plattform in targeting mobile and desktop-computers and it is not common for frameworks to provide source-code, see Table 3.1. While frameworks typically support more than

one modality, only few frameworks support more than three modalities. Similarly only few frameworks provide more than three classifiers. Prototyping support via configuration files or UI is also not typically provided.

MediaPipe, a recent addition to synchronized processing of audio-visual data [114], is open source and runs on various platforms. It's focus lies in video processing and therefore does not support other sensors e.g. for capturing physiology and thus, does not provide according feature sets as well. It has no means of actively engaging in machine learning processes but rather only relies on their outcome.

Conclusively the approach presented in this work (MobileSSI) distinguishes itself from others by spanning mobile as well as desktop platforms and being feature-rich across aspects of integrated sensors, classifiers and flexibility in configuration and UI. It investigates corner cases of MSSP such as environmental contexts and drinking activity, next to the paralinguistic social clue of laughing. On a technical note it tackles privacy via interactive machine learning on mobile devices and uses custom tailored fusion approaches to multimodal data processing. Technical aspects are described in more detail within the next chapter.

Framework	Modalities	Online Classification	iML (online, active, UI)	Prototyping	Emotion Recognition	Privacy	Environmental Context	Plattform	Open Source
MobileSSI [68]	Audio, Physiology, Activity, Environment, Video,..	ANN, SVM, NB, kNN, $1	Y/Y/Y	XML/Web	Y	Y	Y	Windows, Linux, Android, OSX	Y
SSJ [44]	Audio, Physiology, Activity, Video,..	ANN, SVM, NB	Y/Y/N	XML/Flow-Graph	Y	Y	N	Android	Y
CenceMe [130]	Audio, Activity	DT	N	N	N	Y	N	Symbian	N
Jigsaw [113]	Audio, Activity	DT, GMM	N	N	N	N	Y	Symbian, iOS	N
MediaPipe [114]	Video	ANN	N	JAML/Flow-Graph	N	N	N	Android, iOS	Y
SociableSense [158]	Audio, Activity	GMM	N	N	Y/N/N	N	N	Symbian	N
SoundSense [112]	Audio	kNN, GMM, HMM	N	N	N	Y	Y	iOS	N
EmotionSense [157]	Audio, Activity	GMM	Y	N	Y	N	Y	Symbian	N
Auditeur [136]	Audio	DT, NB, GMM, SVM, HMM, kNN	N	XML	N	N	N	Android	N
CRN [12]	Audio, Activity, Physiology	HMM, kNN	N	Flow-Graph	N	N	N	iOS, Symbian	Y
EEMSS [224]	Audio, Activity	DT	N	XML	N	N	N	Symbian	N
FieldStream [61]	Activity, Physiology	SVM	N	N	N	N	N	Android	Y
BeTelGeuse [103]	Audio, Physiology, Activity, Phone Usage, Video,..	SVM	N	N	N	N	Y	Windows, Linux, MIDO	Y

Table 3.1: Mobile Frameworks for continuous signal processing and classification.

Chapter 4.

MobileSSI - Framework and Implementation

MobileSSI is no different code base from the Social Signal Interpretation framework (SSI), but rather an effort to bring the code base to mobile smart devices. Therefore, an introduction to the approach to Social Signal Processing (SSP), as implemented in the SSI Framework, is given at the beginning of this chapter. This chapter summarizes the technical aspects added over the course of my publications [65–69] and goes beyond in showing own work on user-interfaces (head ache diary), active learning (using SVMs) not published. Since SSI is a collaborative effort it is hard to draw clear lines, where one's contribution ends and another one's starts. Yet, developing SSI for mobile and embedded devices is an own contribution to the software.

SSI - the Social Signal Interpretation framework is designed as a toolkit to deliver input from natural social behavior to other applications. Thereby, it provides real-time feedback via sensing (Section 4.3), processing (Section 4.4) and classification, (Section 4.5) that is communicated to further application components (Section 4.8). To deliver classifiers for custom applications, SSI also provides features necessary for recording, e.g. stimuli presentation and machine learning, e.g. model evaluation.

This chapter also contains an description of the process of porting SSI to UNIX-systems and Android in Section 4.2. Even though MobileSSI is part of SSI, it includes contributions that are not useful, when running SSI on other systems than Android e.g. integration with the Android operating system. Components and approaches in signal processing, machine learning, recording and communication developed for the mobile use, form the main part of this chapter.

There are a number of frameworks that specialize in processing social signals on mobile devices, such as SociableSense[158]. This framework has the advantage of implementing data sampling strategies that are more suitable for mobile devices. Porting SSI as an established framework has the advantage of tested feature-sets and machine learning capabilities. Considering Table

3.1 only BeTelGeuse [103] and SSJ have an amount of sensors, comparable to MobileSSI. BeTelGeuse has the limitation of not running on current mobile platforms, such as Android whereof SSJ cannot be used on desktop machines or on embedded devices. Media-Pipe [114] is a recent addition to mobile processing frameworks that is limited to audio-visual data only and is therefore not suitable for the purpose of a companion on the body that relies on accelerometer, audio and physiological data. The requirement to have a framework with wide spectrum of functionality that can be tailored to the needs "in the wild", motivated the decision to develop MobileSSI. SSI, also including the mobile port is open source and available on Github [1] together with an Android Studio project for building the mobile application [2].

4.1 Continuous Processing and Synchronization

SSI is a modular system that can be configured to solve different problems such as the recognition of gestures, postures or laughter [16, 116], loading different plugins at start-up. A core problem SSI is designed to solve is recording multiple continuous signals of a fixed sample-rate, which are synchronized. The sample-rate is defined before start-up and watched over each sync-interval to keep signals from drifting. In the case of a sensor providing to few data, data is added. For sensors providing too many data, data is removed. There are different approaches available, e.g. repeating data in the case of a slow sensor or filling with zeros. These continuous signal *streams* are processed with filters and can be fed into classification and machine learning processes. Streams are of a fixed type and can be multi-dimensional.

4.1.1 Events

Beneath continuous data streams, there exist events in SSI, that are sporadic and thus come with a time-stamp and duration. They are rather designed for flow-control triggering feature calculation or classification, then information processing and therefore can not directly be learned or classified. Events can contain complex information formatted in maps and strings and as such high-level information. This information is usually the interpretation gathered from low-level data via classification and is forwarded to other applications finally.

4.1.2 Processing Pipelines

To gain information of a higher level, processing in several stages and lines is employed. The defined set of those components is called a pipeline. Based on the flow from a source (*sensor*) through a processing unit (*transformer*) to a final stage of the flow (*consumer*), the basic types of

[1] https://github.com/hcmlab/SSI
[2] https://github.com/hcmlab/mobileSSI-android-studio

SSI components can be identified. The sensor can provide multiple output-signals of different type and sample rate through *channels*. A basic entity with the ability to receive and send events is called *object* and forms the basic class in the inheritance hierarchy. The pipelines construction

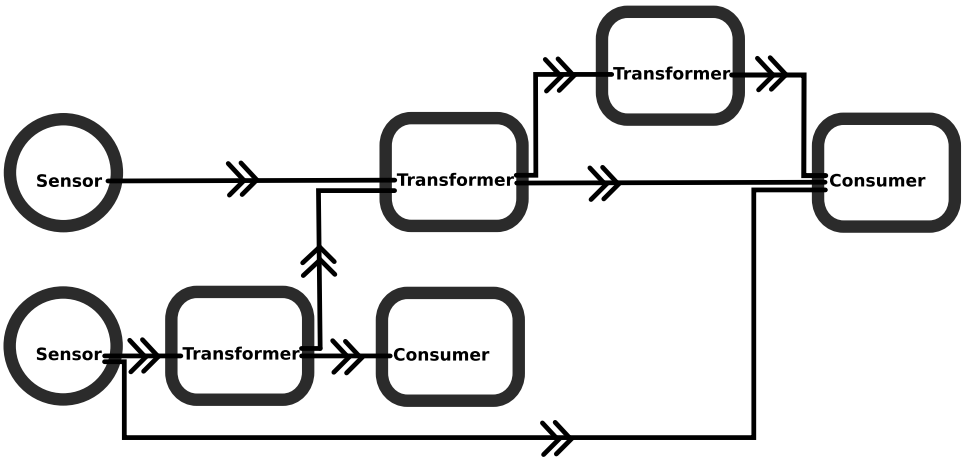


Figure 4.1: SSI pipelines consist of three basic types: sensors, transformers and consumers.

can happen via C++ code, but typically is done using XML. Parsing is done using the XML-Pipe component that triggers the *Factory* found in *core* to load library and *theFramework*, found in the *frame* plugin, for assembling and starting the pipeline.

4.1.3 Soft Real Time Processing

The organization in processing pipelines is important for real time recognition. A sample can be processed while a new one is recorded simultaneously. The sample consists of a time-window, also called frame-step, of certain length (e.g. 400 ms) and can be accompanied by data that overlap (e.g. 1 s) into the past (delta) or ”future”, resulting in an overall duration of 1400 ms and with a sample-rate of 5 Hz into seven actual values. In SSI those overlapping data are called left (past) of right (future) context.

SSI does not enforce real-time behavior in strict time constraints. Plugins for pre-processing and classification as well as window-sizes and system have to be picked carefully to achieve a responsive system, since lag might add up over time.

4.2 Mobile Port of SSI

SSI is mainly written on and for Microsoft Windows machines using Visual Studio and C++. It uses the Win32 threading API, timers and event system. Visual Studio provides both the build system and the compiler. After contributing additional code paths to the existing framework,

plugins and concepts are added to make it a better fit for the new circumstances on mobile platforms

Porting to Mobile

To adapt the code base for mobile devices, a different compiler and build system is needed, as well as changes to the core system for threading and timers. This results into four steps:

1. replacing the build-system and compiler
2. replacing the Win32 threading model with C++11/Posix
3. changing the operating system to Linux
4. building for Android

At the beginning of 2015 there was no native build support in AndroidStudio, but with a set of CMake files it was possible to not just build the code but also APK-packages for Android. Since CMake has good cross-platform support on Windows and Linux, it was the build-system of choice.

CMake has Visual Studio support, which made it possible to test the new build-system with the old compiler, before moving to MinGW. MinGW supports Win32 as well as Posix threading, thus it is possible to replace the threading model without changing the OS.

The change to Linux consisted mainly in adapting the dynamic library loading, timers to `clock_gettime()` and `CLOCK_MONOTONIC_RAW` and rewriting the event system using condition variables. The Gnu Compile Collection was the natural choice after using its Windows derivate MinGW.

To be able to run SSI's UI based tests, a new GUI-backend was written using Cairo and SDL2. The code base was tested and fixed to run on ARM based platforms such as the RaspberryPi as well.

Having SSI ported to Linux, building the core libraries for Android using the NDK's cross-compilation toolchain and corresponding CMake scripts was the next step.

This process of porting marks the mere foundation on which adaptions of the core plugins such as audio and additional plugins, e.g. for accelerometer are, built. Since SSI is now a portable native library, it runs on further mobile platforms and has been used on Tizen Watches such as the Samsung Gear S2.

Android Integration

While first attempts were made to build MobileSSI into a native App, using the Andoid-SDK to load native libraries via Java Native Interface (JNI) turned out to be the more viable approach. SSI here is started as a background-service that is using a wake lock to be able to continuously record data. Since various devices such as GPS and bluetooth devices are hard to integrate using native APIs, it is possible to send data via JNI to MoblieSSI's native libraries, e.g. for recording.

SSJ Integration

In addition to the native port of SSI there is SSJ, an effort by Damian et al. [44] to rewrite SSI in Java to run it on Android devices. It allows easy access to sensors via Android's Java-API and the design of processing pipelines via a flow-graph based GUI. There are two ways of integrating SSJ and MobileSSI that supplement each other. SSJ can function as a sensor to MobileSSI and conclusively a MobileSSI consumer is integrated into SSJ pipelines.

MobileSSI plugins can be loaded in SSJ using a wrapper instead of running the framework as a whole. This enables SSJ to run compute-intensive filters and feature extractors in native code.

4.3 Sensors

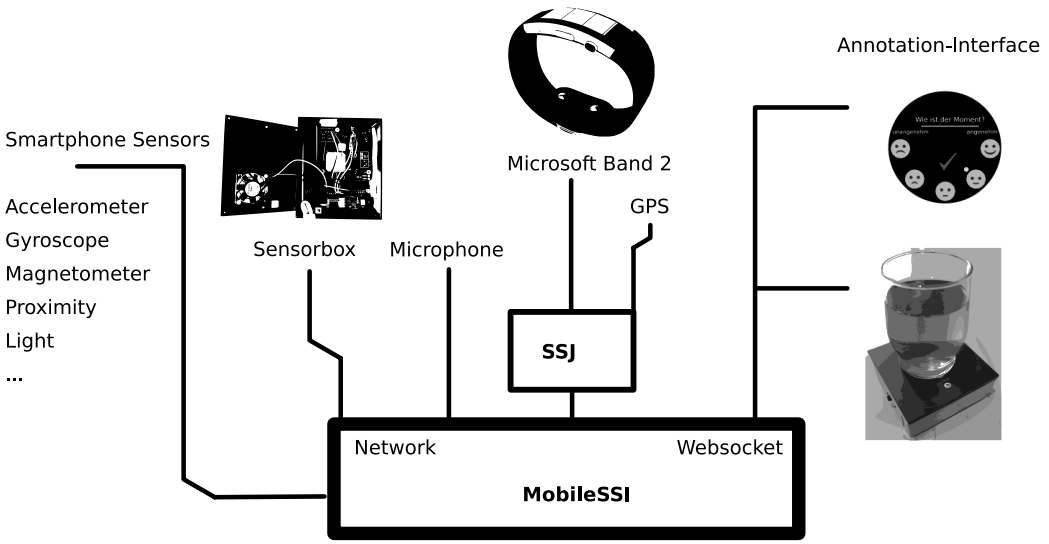


Figure 4.2: Sensors integrated into MobileSSI.

MSSP is about extraction information from natural communication, that typically needs sensors converting analogue signals into digital data-streams. As mentioned in Section 3.6 there is a

shift in perspective from desktop or lab SSP to MSSP "in the wild". This made it necessary to integrate new sensors into SSI next to adapting existing implementations. An overview is given in Figure 4.2.

Audio depicts the bridge where an old code base is extended on the new platforms. The smart phone sensors are specific to Android but accessible via native C-API, while environment sensors are connected via a (UDP-)network, the physiological and GPS data via SSJ and the annotation interfaces via Websocket-network (smart watch) or Bluetooth (smart scale).

4.3.1 Audio

Audio-support is realized using the Win32-API on Windows. Next to the camera-plugin it forms the core modalities on the desktop. Audio on Linux is implemented using port-audio, that is not available on Android, where the OpenSL-API is used instead.

The focus lies in recording audio data at a variety of sample-rates and formats. While on PC sample-rates of 44kHz and higher are common, on smart phones audio-sample-rates often are limited to 16 kHz. Also Android does not always provide data at the requested sample-rate and one might actually record at 8 kHz instead of the requested 16 kHz.

Since streams in SSI have strict types and many filters rely on floating-point types for processing, while the hardware does provide 16-bit integer data, the data is converted within the audio-plugin when option *scale* is set.

The audio-plugin contributes a range of features next to recording capabilities. On Windows and Linux there is playback support for audio files also. There are different approaches to Voice Activity Detection (VAD) contained in the plugin, that will be described under Section 4.4.2.

4.3.2 Accelerometer and Android-Sensors

Motion is a core modality in MSSP. While motion would be extracted from video data in the lab, inertial sensors are used in mobile set-ups. Accelerometer for linear acceleration and Gyroscope for rotation acceleration are provided on all modern smart devices. With different platforms there is a need for different APIs to be integrated. On Tizen therefore a different plugin is created instead of using one plugin with different code paths. Since Android-Sensors have to be enabled and requested in an initialization process, all smart phone sensors connected via the Android-C-API are combined in one plugin. Since sample-rates might differ, multiple streams are provided as the plugin's out-put. Those streams can have multiple dimensions. This is the case for accelerometers that have three axis.

Since the "AndroidSensors" plugin handles the core functionality of SSI on this platform, the functionality to execute XML-pipelines is part of this plugin as well.

Android sensors might not be available to the full extent on all smart devices, the maximum modalities supported in MobileSSI are:

- Accelerometer
- Gyroscope
- Magnetometer
- Light
- Proximity
- Heart-Rate (only on smart watches with BVP-Sensor such as the Moto360)

Further sensors only available through Android's Java-API, such as GPS, are realized using the "AndroidJavaSensors" plugin in MobileSSI with the according Java functions provided via JNI.

4.3.3 Physiological Signals

While some physiological sensors are built into the smart devices themselves, formost in smart watches like the Motorola Moto360 or the Samsung Gear S2, they are mostly external devices connected via Bluetooth. The key feature of MobileSSI here is to access the data-source directly, without the use of cloud services.

This need for access to raw data has two reasons. Firstly, it is desirable that the user has a choice with whom to share privacy-critical data. Secondly, MobileSSI has to synchronize the physiological sensor with other modalities, such as sound and process the data within a window of a few seconds to deliver feedback reactivity.

As external sensor for physiological signals the Microsoft Band 2 is used in Chapter 7. It provides heart rate (HR) and inter beat interval (IBI) via a Bluetooth-based API. Mobile Sensors mostly rely on Blood Volume Pressure (BVP), an optical method based on green or blue LEDs and on the amount of light absorbed by body supplied with blood. Moto360 and the Microsoft Band 2 provide data that is already processed, whereas the Samsung Gear S2 allows access the raw sensor-data.

Since MobileSSI is able to process raw sensor data, low-level data access means a broader control and a wider set of features that can be calculated.

The Microsoft Band 2 is integrated into a MobileSSI pipeline via SSJ. It provides skin conductance (SC) next to BVP as a physiological signal. Here the resistance of skin between two electrodes is measured. Usually under stress, physical or mental load, SC increases in the form of spikes, but is delayed from the source event.

4.3.4 Environment Sensors

Physiological sensors are worn close to the body, whereas environment-senors take an outwards-perspective on a situation's context, that is rarely used in the lab. Anyone who has ever gone for a walk along a busy road and in a forest on a hot summer day can tell the difference an environment can make to the same activity.

Temperature, humidity and air-pollution are the three modalities MobileSSI is able to sense using its environmental sensors.

Device	Sensor	Data	SR (Hz)
Sensor Box	SDS011	PM2.5	0.07
		PM10	0.07
	SHT75	humidity	0.07
		temperature	0.07
	MICS	CO	0.07
		NO_2	0.07
		NH_3	0.07
		C_3H_8	0.07
		C_4H_{10}	0.07
		CH_4	0.07
		H_2	0.07
		C_2H_5OH	0.07
	BMP280	pressure	0.07
		temperature	0.07

Table 4.1: Environment Sensors integrated into MobileSSI via UDP-Network.

As Table 4.1 shows, temperature is measured by SHT75 and BMP280 sensors, whereas SHT75 also measures humidity and BMP280 can also measure air pressure.

MICS sensors are used to detect a wide range of gases, see Table 4.1, such as carbon monoxide and ammonium. The fine dust concentration is detected via SDS011.

Those gas sensors are not calibrated and as such can not give an absolute concentration but rather a trend. The sensors' behavior might change over time and is dependent on other factors e.g. humidity.

Similarly, noise pollution could be captured with measuring microphones that have undergone a calibration process. While a microphone could be used via USB and the audio plugin of MobileSSI, the sensors are built into a custom build sensor-box. The box has a Raspberry Pi Zero, where the sensors of Table 4.1 are connected by wire and serial bus. The smart phone hosts an ad-hoc WIFI network over which the data are sent as multidimensional stream via UDP to the MobileSSI instance running on the phone.

4.4 Features

Within the flow of the processing pipeline, the sensor recording is followed by feature extraction Feature extraction marks aspects of a input signal that are of relevance and helpful for the classification process. This ranges from pre-processing the data in such a way that the classification can be as simple as applying a threshold. However, it can also be integrated into a deep learning process in which an artificial neural network with multiple layers learns features in its layers that are closer to the input with increasing abstraction towards the output.

Statistical Functionals

A generic form of features looks at the data statistically. What is the mean, minimum or standard deviation of the data? A set of statistic features is sometimes referred to as functionals in the literature [183] and also in SSI. Functionals are useful as an additional pass (long-term) over a big set of features or as a base from which to add specific features. SSI's functionals consist of twelve individual values, whereof the following nine are used e.g. as starting point of accelerometer-features in Chapter 5.

- Mean, Standard deviation
- Minimum, Maximum, Range
- Zero crossing rate
- Peak count, Pulse rate
- Energy

Time Domain

Features can work on the data as recorded, a sequence of samples gathered from a sensor. This time-domain-features can be based on functionals, e.g. calculating integrals. Peak-detection can also be implemented in time-domain. Wavelets [151] reconstruct the original signal by iteratively adding scaled and translated variants of the mother wavelet curve. This can be used to identify a signals characteristics.

The reconstruction is called (Fast) Wavelet Transform and as such describes a way of identifying frequency spectrums of relevance in a data-class, even though they are applied in time domain.

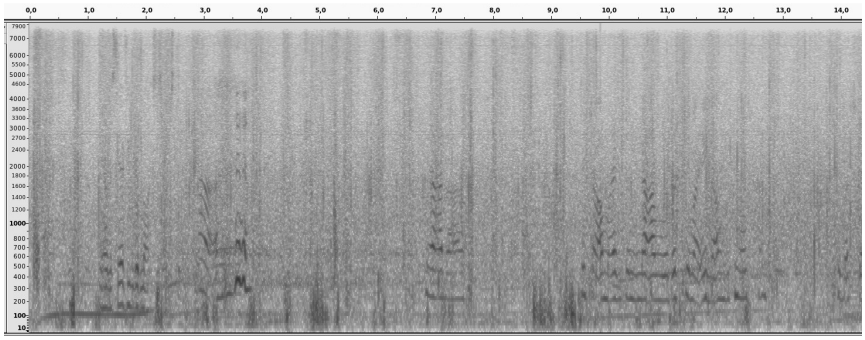


Figure 4.3: Spectrogram of human voice (dark waves, e.g found between 1.0 and 2.0) and background noise (grey noise pattern throughout the recording), Mel scale. Taken from the data set of Chapter 7.

Spectral Domain

Separating spectral and time domain features comes from the fact, that many features require a (Fast) Fourier Transformation (FFT). As a result of an FFT, data are organized along their spectral property. The FFT calculation can be done once on the raw signal and shared across all spectral domain features.

In spectral domain, it is easier to focus on a certain band of frequencies or determining the most dominant frequency, see Figure 4.3, where the phonemes of human voice are clearly recognizable (dark wave pattern) even tough there is a lot of background noise (grey noise pattern across the picture). In an ECG signal this would be the heart rate. Spectral domain features are used in a big variety of modalities, next to physiological signals such as Galvanic Skin Conductance or Blood Volume Pressure, they are used in accelerometer and audio data.

Feature Reduction

An option to reduce compute resources in feature calculation is feature reduction. A common method of feature reduction is Sequential Feature Search (SFS). It determines the single best feature and to that adds the feature that improves the classification result the most (or worsens the least). This way features are ranked and one can restrict the feature-set to those giving the best results. SFS is used to give insights on the contribution of individual heart rate related features in Chapter 7. This feature reduction can lead to over-fitting and therefore the resulting model may perform worse on new data.

4.4.1 Acceleration

Accelerometer features typically are based on statistical features that are applied on the individual axis, but also considering multiple axis. Those functionals involve: Mean, Standard

Deviation, Energy, Root Mean Square, Variance and Interquartile Range. Additional time domain features are Haar-Filter and a Discrete Cosine Transform (DCT).

In spectral domain Entropy, Flux, Centroid, FFT-Sum and Rolloff are computed. Accelerometer Features in Mobile SSI are based upon work by Dietz et al. [49] who used the features for head movement analysis. The feature set was extended to fit drink activity recognition and is used in Chapter 6. Features can be computed normalized between 0 and 1 to be used without normalization in an additional step in incremental training.

4.4.2 Audio

Audio feature sets are usually in the spectral domain, after applying a Fast Fourier Transform (FFT). This is true for audio-object recognition as well, see Section 4.4.2. For speech and paraliguistic processing there is a Mel-Scale applied for the spectral domain, which is than aggregated in Mel Frequencies Cepstral Coefficients MFCC [105, 235]. MFCCs simulate human perception, and generalize a signal in a way that phonemes are underlined. While MFCC's by themselves might be calculated over short time windows of ~40ms, there are feature-sets specifically designed for audio (emotion) recognition that also consider longer time frames and are used in Chapter 5.

EmoVoice

EmoVoice [219] library provides a feature set (V2) consisting of 1451 features combining pitch, energy, MFCCs, Frequency Spectrum Quantile and Harmonics-to-Noise Ratio (HNR). Pitch calculations are based on Praat, a tool for speech processing and phoneme-annotation [26], that works offline. Although EmoVoice, like OpenSmile, provides tools for training and execution of emotion recognition, they are only considered as libraries providing a feature-set in this work.

OpenSmile

OpenSmile [63] provides different feature-sets, like *ComParE* [128] with 6373 features and a minimal *GeMAPS* feature-set with 58 [62] features. OpenSmile also considers MFCCs, Pitch, Harmony to Noise Ratio (HNR) as well as functionals over those low-level descriptors (LLD).

Audio Activity Detection

Large feature sets are expensive to compute. Within a frame based approach feature calculation can be paused as long as no Audio Activity (AAD) or Voice Activity (VAD) is detected. SSI's

audio plugin has different approaches to detecting audio activity. The naive approaches are loudness or intensity based, while more sophisticated approaches are based on signal-to-noise ration (SNR). Further approaches provide own models to solve the problem of VAD.

Features from Deep Learning

While deep-learning per se is a step towards feature-less machine learning models, where training happens directly on raw data, the layers closer to the input can be cut from the classification layer (closest to the output) and used as a feature-set themselves. Thus, other classifiers such as SVMs can be trained on deep-features, to speed up the training process. See Section 4.5.4 for details. This method, although generic and applicable to many different problems, is used in Chapter 7.3.3 to use an extractor for image features, based on MobileNet V2, on spectral maps of audio data.

4.5 Classifiers and Learning Approaches

At the core of Social Signal Interpretation is the extraction of abstract information. Nowadays this step is mainly realized by machine learning. There are several procedures and structures involved in the process, the most important procedures are training, evaluation and prediction. Those can be abstracted to support multiple learners, e.g. k-fold cross validation is implemented independent of the model implementation in (Mobile)SSI. The individual learning algorithms or classifiers still are the most central structure. This section takes a closer look at classifiers integrated and used with MobileSSI.

A modification in MobileSSI regarding the learning process involves users and thus makes machine learning more interactive: *reactive* and *active*. Reactive, while predicting not on a whole recording, but on a short time frame, to give useful feedback in real time.

Active, as in the model participating in the learning process by proactively asking for labels. The model becomes reactive also in a sense of shorter time to perceived change in the model. Usually learning happens in batches of samples, it is desired to use small batches, or even individual samples to update a model incrementally.

Bigger batches enable pre-processing, such as normalization or model specific parameter evaluation to make learning more efficient. Real world data would have to be aggregated into those batches, which requires time and also storage of potentially critical data. Reducing the batch-size foremost makes the learning system more responsive, see Figure 4.4. This process leads to stream based, incremental or online learning.

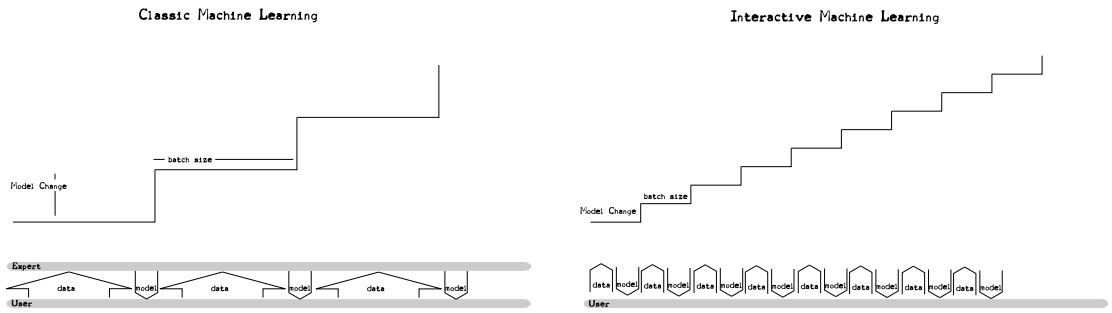


Figure 4.4: Roles and updates in classical and interactive machine learning.

Balancing

Many ML algorithms (SVM, ANN) are sensitive to imbalanced distribution of samples per class [15]. This goes as far as incremental learning on a batch with a subset of classes results in a new model that supports only that subset of classes. Aggregating samples takes, dependent on the problem at hand, a considerable amount of time, balancing mini-batches with older samples might make the training harder to grasp for the user. Thus, interactive learning has to come up with a pooling approach.

4.5.1 Query Methods in Active Learning

When creating an ML-Process that involves both model and user as active parts, selecting the right samples to ask the user for labels is crucial.

Active Learning [192] provides several techniques for asking an ”expensive oracle” for information. This oracle can be a compute intensive simulation or a human, in the case of this work human are the only considered oracle. They can involve general approaches, approaches involving multiple models and model-specific methods. Model specific approaches for SVMs are outlined at the introduction of SVMs later on. *Uncertainty Sampling*, applicable to different models and *Query by Committee* involving multiple modes are introduced in the following.

Uncertainty Sampling

Samples can be selected for training by taking the existing model's certainty in classifying that new sample into account. Studies [108] led to the insight, that samples classified with low certainty speed up the learning process more compared to selecting only samples with high certainty regarding their classification. If the classifiers in use are able to provide certainties, adding a threshold is sufficient to filter samples according to their usefulness.

Query by Committee

Another method to generate requests is to use multiple models [193]. Classification results are viewed as vote for the according classes. Here a common rule is to ask the oracle, the more likely, the more diverse the voting is. The committee of classifiers is created to be as diverse as possible by itself [127].

4.5.2 Naive Bayes

A simple but still successful approach to classification is Naive Bayes. Here the Bayes formula on conditional probabilities is the foundation.

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Where a conditional random event $P(A|B)$ can be expressed by its inverse $P(B|A)$ in relation to the unconditional probabilities of properties A and B. With respect to classes C_i and features $x_1..x_n$, we would like to know the probability of a class C_i given the observed features:

$$p(C_i|x_1, ..., x_n)$$

Which with application of the Bayes formula results in:

$$p(C_i|x) = \frac{p(C_i)p(x|C_i)}{p(x)}$$

With the a-priori probability $p(C_i)$ the conditional probability of a feature x given a class $p(x|C_i)$ and the probability of that feature x $p(x)$. The features are naively assumed to be mutually independent thus, the assignment of the class can take place via:

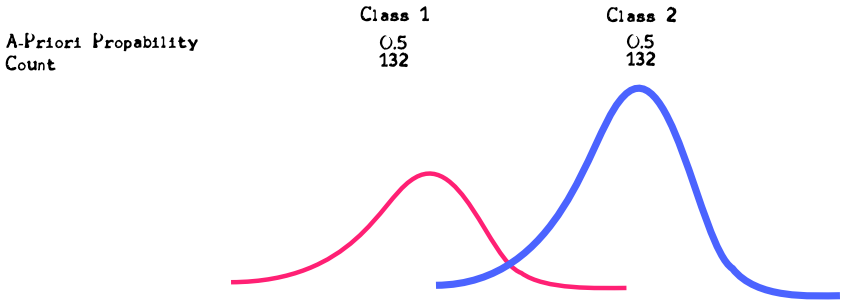
$$c_i = argmax_{i \in \{1,..,I\}} p(C_i) \prod_{j=1}^{n} p(x_j|C_i)$$

Given the circumstances of feature calculation, depending on a single sensor of at least a single matter, this assumption is nearly always false. Nonetheless, Naive Bayes gives an approximation that is useful in many cases.

For continuous variables (integer of floating-point) Gaussian distributions are common, and $p(x|C_i)$ for an observed value v can be expressed using mean μ and variance σ of the feature x in he training set.

$$p(x = v|C_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{(v-\mu_i)^2}{2\sigma_i^2}}$$

Naive Bayes models, as found in MobileSSI, consist of tables storing mean and variance as well as standard deviation (σ^2) and a-priori Probability ($p(C_i)$) of each class C_i (see Figure 4.5).



	Standard Deviation	Variance	Mean	Standard Deviation	Variance	Mean
Feature 1	0.005382	0.004252	0.005671	0.005836	0.003324	0.005018
Feature 2	-0.001335	0.000877	0.002578	-0.000762	0.000712	0.002323
Feature 3	-0.011234	0.010383	0.008869	-0.011128	0.009465	0.008468

Figure 4.5: Schematic of Naive Bayes data structure for online learning [77].

Incremental Learning Naive Bayes

This can be adopted to per sample learning by adding a sample counter and adjusting mean and standard deviation incrementally [77].

$$\mu_n = \mu_{n-1} + \frac{x_n - \mu_{n-1}}{n}$$

$$\sigma_n^2 = \sigma_{n-1}^2 + (x_n - \mu_{n-1})(x_n - \mu_n)$$

The principle of adjusting mean and standard deviation is not limited to adding one sample, but could be extended to merge two models. This makes Naive Bayes a good test bed for different approaches.

Naive Bayes per se delivers certainties $p(C_i|x)$, to the degree to which it believes the model describes the situation accurately in his prediction, which enables querying on uncertainty.

4.5.3 Support Vector Machines (SVM)

Support Vector Machines (SVMs) separate data of different classes by finding an according *hyperplane*, meanwhile observed samples are used as *support vectors* to describe said hyperplane.

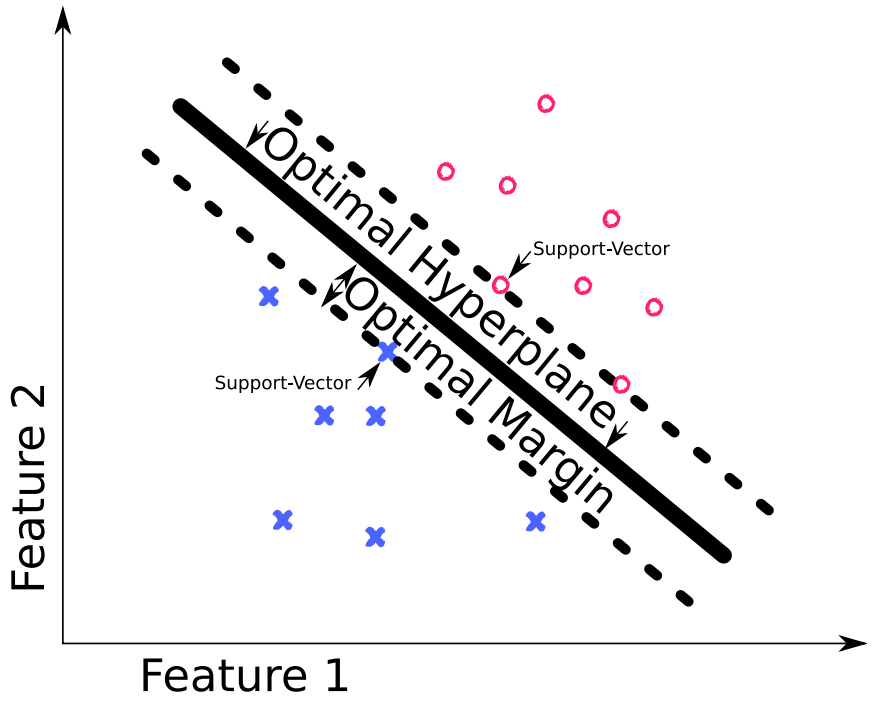


Figure 4.6: Schematic of a linear SVM [38].

As the hyperplane separates the support vectors, there is a space called *margin* between support vectors, with multiple possible solutions. The smaller the margin, the better defined the hyperplane is.

Viewed from the stance of SVM, samples are located in a space with one dimension per feature. The simplest form of SVM operates with a linear kernel, where feature space and kernel space are the same. More advanced kernels are often available that transform the samples in a way that makes them linearly separable, this thesis restricts itself to linear SVMs. Next to the applied kernels there are further simplifications made in Figure 4.6, such as the introduction of soft margins. Soft margins allow balancing margin-size and outliers in the case of not linearly separable data.

Platt Scaling

SVMs cannot per se judge with what certainty an observed sample is put into the according class. To extend SVMs' capabilities, Platt-Scaling was introduced [152]. It is a logistic transformation, that considers the samples of the training-set to build probability distributions. Since the iteration over all training samples is needed, Platt-Scaling hinders SVMs from online-learning.

Incremental Learning SVM

Nonetheless, online and incremental learning approaches for SVMs exist e.g. in LibLinear [210] and DLib [101, 196]. LibLinears approach builds upon a warm-start model, small batches can be used to incrementally improve the model. Those batches should be as big as possible and have to contain samples of all classes, or else the incremental learning leads to a model with reduced class-count.

Hyperplane Based Queries

Next to general query methods, there are model dependent ones. In the case of SVM's it is close at hand that queries are derived from the hyperplane that is essential to SVMs' classification process [208]. A possible method is to select samples that possibly half the margin and thus reduce the space in which hyperplanes could lay in. This is of interest in a case where general methods such as query on certainty fail because of missing confidences in online learning SVMs or if an additional contribution to a committee is desired.

4.5.4 Artificial Neural Networks via TensorFlow

Artificial neural networks are supported in SSI in two ways. TensorFlow is integrated via a Python interface for training on one hand and via C-API for mobile use on the other. Python on Android is not supported to such extent, that it allows running full TensorFlow for training. The C and C++ APIs do not support the training process to an extent suitable for our purpose. This allows MobileSSI on Android to only load models for classification via a TensorFlow-plugin. An approach allowing for DNN-training on Android smart phones is running a Gnu/Linux via Chroot [188].

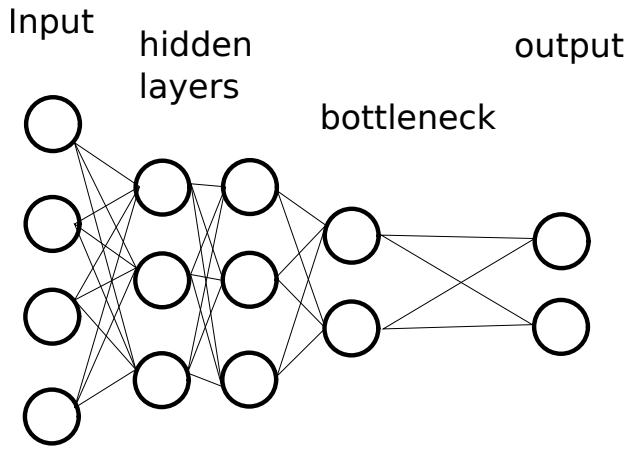


Figure 4.7: Schematic of an ANN.

Deep Artificial Neural Networks, as common today are built of a multitude of components. For simplification and due to the fact, that this thesis only requires a view on a certain abstraction, only a rough overview is given. DNNs consist of nodes representing a Perceptron or a comparable more complex structure and weighted connections in between. For classification, the network starts at input derived from a sensor that is fed into hidden layers and lead to the output layer representing the classes, the network targets. The layer before the output can be seen as feature extractor of the network and is called bottleneck in the description of current network architectures such as MobileNet v2 [180].

4.5.5 Transfer Learning

The approach described in "Freatures from Deep Learning", technically is a form of (transductive) transfer learning, *"where knowledge from one domain is used to solve a problem from another domain"* [144].

In compute-intensive training of Deep Neural Networks this allows to train a model within minutes on a smart phone based of an Artificial Neural Network (ANN) that took days to train on a server-farm. Next to compute power, savings in the size of the data set makes a difference. While millions of images are needed to train e.g. ImageNet [45], only hundreds are needed to do re-training [188].

Instead of training a full network, only the last layer responsible for the output is replaced with a layer for the new target classes.

Re-training takes place also in warm-started incremental learning (e.g. using SVMs), where the number of classes of the new model does not have to match the number of classes of the old model.

4.6 Fusion

A core challenge in SSP is using multiple modalities to solve the sensing problem to create a more accurate and reliable system. The process of gathering joined information from two or more signals is called *fusion*. Fusion can be handled at different stages of a processing pipeline and considering a variety of strategies (synchronized/asynchronous) and algorithms. A wide range of approaches is available in SSI's Fusion and Vector-Fusion plugins, that build also on Android.

Feature Based Fusion

Feature based or early fusion happens before data are fed into a machine learning model. Thus, it handles the combination of features. In online recognition the fusion of feature vectors works only when the data are processed in a sliding window of the same size. On merged feature vectors a single classifier for all modalities can be trained, which allows e.g. feature selection across modalities to identify the share and aspect they contribute to the machine learning solution.

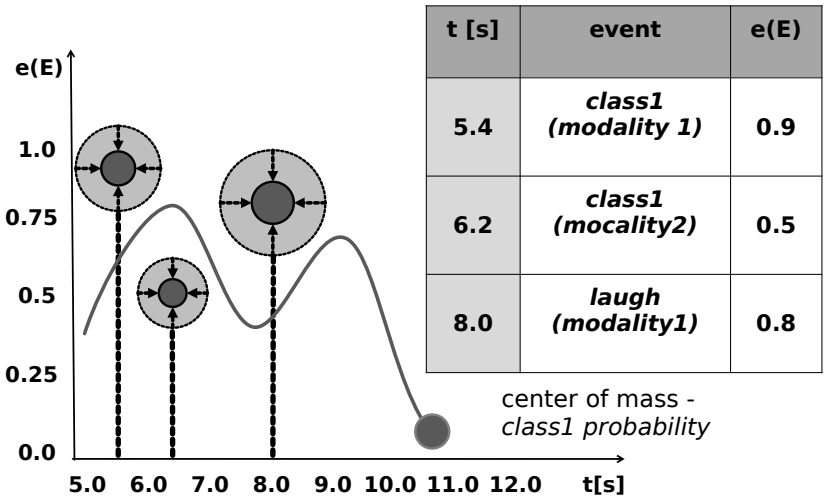
Decision Based Fusion

Decision based fusion, also called late fusion is happening after the training of individual models per modality, that conclusively handles the combination of different models' results. This can happen by having the ensemble of classifiers vote, adding or multiplying their certainties or following a similar rule. While not used in the scope of this thesis boosting is also a form of late fusion, where an ensemble of weak classifiers are combined into one strong classifier [75].

Asynchronous Fusion

Feature level fusion is bound to having features work on the same time step or window. Individual modalities convey meaningful information in segments of different length. To cope efficiently with changing sources of information that depend on sensors available and information emitted by the surrounding, fusion on the event level is employed. This asynchronous fusion approach does not force decisions from all available channels for every time frame, but instead correlates occurrences of small windows of relevant information over time. Other ways of fusing modalities without steadily forcing decisions have been successfully investigated in other approaches: Zeng et al. [236] apply Multi-stream Fused Hidden Markov Models, in which state transitions of different components of Hidden Markov Models are allowed to occur at differing times across multiple streams. Dupont et al. [53] model the asynchronous nature of audio and video streams using temporal typologies with multi-stream Hidden Markov Models for continuous speech recognition. Methods pursuing a hand modeled approach for the asynchronous fusion of streams using Petri-Nets are applied by Navarre et al. [135]. Whereas fusion engines often focus on high-level dialogues, this chapter's focus lies on a lower level and is supported by learned models and events solely. Long Short-Term Memory Neural Networks have shown great success in paralinguistic tasks (see for example [29, 228]).

They were used to replace simple nodes with memory cells that allow the network to learn when to store or relate to bimodal information over long periods of time. Asynchronous fusion on event level has proven to be robust in affect recognition scenarios [110]. Events are interpreted



t [s]	event	e(E)
5.4	***class1 (modality 1)***	0.9
6.2	***class1 (mocality2)***	0.5
8.0	***laugh (modality1)***	0.8

Figure 4.8: The fusion algorithm considers the temporal flow of occurring, *class1* indicating events. Influence of events decreases over time and a continuous probability of *class1* is calculated as the moving centre of mass of weighted events. *Courtesy: Florian Lingenfelser*

as short-termed cues that point to a searched target class. An ensemble of trained machine learning models is used to recognize these events in the available modalities. By monitoring occurrences of events over time, the higher level fusion plugin is able to decide what is going on at any point in time. This strategy provides an abstraction level that allows for easy adaption, as modalities that are able to provide events for the event based fusion algorithm can be easily added or removed. Therefore it is a good fit for ”in the wild” signal processing, where it is not guaranteed to have all sensors available at all times (differing hardware, noisy data, energy consumption, etc.). Recognized events are initially weighted with regard to the confidence of the classification model and this weight is constantly decreased so that their influence on the final fusion result descends to zero over time.

Currently active events give an appropriate overall picture, which in turn is judged by the fusion model on the event level by calculating the center of mass based on currently active events and their updated weights, see Figure 4.8. On one hand this solves problems with different sampling rates and segmentation windows on different streams. On the other hand, it is especially useful if the events of interest can be expressed differently in multiple modalities. From an engineering point of view, only transmitting events is leading to much lower network load than sending a raw data stream.

4.7 Interactive Machine Learning

In classical machine learning, the process of model creation is hidden from the user. A expert solves data extraction, processing, labeling and training. The resulting ”AI” solution is deployed to the user. Next to user control in model creation, additional methods of feature creation and combination as well as explanation are proclaimed by Amershi [5]. The focus within the implementation of MobileSSI is on creating a solution that is reactive, as in online learning, pro-active, as in active learning and easy on resources, as in transfer learning, so that the user can be involved in the training process.

1. This is realized by an active learning model that creates label requests based on the certainty of its prediction.
2. The request is forwarded to the user, that labels the sample.
3. With the labeled sample, the online learner is updated.

This process was been refined in MobileSSI’s stream-based process, where a user-given label can be used over multiple subsequent samples according to annotated activity.

4.8 Recording and Communication

For creating a data set, the obvious requirement lies in writing sensor data into a file or a database. MobileSSI as well as SSJ write files to a ”record” folder, that is moved into a folder named after the time-stamp. Recording is done by pipelines that consist only of sensors and the according file-writers as consumers. As the recording process is often a complex operation involving a study design with the presentation of stimuli or at least the generation of annotations, SSI provides a set of tools to support the recording process.

File Writers

First of all File Writers exist for various formats within SSI.

- Streams – Continuous data in multiple dimensions, synchronized
- WAV – Audio files
- Events – Sporadic data points with a time-stamp
- Annotations – Labels in discrete, continuous or free-form

The labeling of recorded data might occur later using tools such as Nova or Elan, but also while recording using user input, e.g. using a web based presentation.

Stimuli-Plugin

Originating in a controlled lab setting is the stimuli plugin, which processes lists of stimuli, iterates them, randomizes them as an experimental setup and writes annotations according to the selection of the stimuli.

While originally combined with a browser plugin, the adapted stimuli plugin iterates over a set of HTML-stimuli contained in a folder. The URL of a stimuli URL is send to the responsible component via event, that might be a websocket instance, described later.

For example a prove of concept of MobileSSI was realized using websites following the Velten method of emotion inducement [215, 219].

Figure 4.9: Velten Stimulus presented via Stimuli-Plugin

The stimuli plugin can be used with more advanced HTML5 presentations such as videos and interactive pages for annotation, such as the Geneva Wheel of Emotion in Section 2.5.1 and Figure 2.2.

Since SSI is specialized in input, it often forms just a part of a larger application. The individual parts have to communicate with each other, what can be solved with MobileSSI in several ways.

App Integration via JNI

On Android, Java or languages based on the JVM are the best supported language for writing apps. In this way, the native set of libraries that form MobileSSI is integrated into a Java user interface to start the background service that runs the SSI pipeline. This is realized on the one hand by loading the C++ libraries in Java and exposing function such as start, stop and communication via events in JNI (Java Native Interface) notation. This allows the Java application to call exposed C++ functions, that can be called API. Native integration via APIs is efficient, since in the best case no copy has to be made of the data processed in different parts of the application. The downside is a loss in flexibility when it comes to programming languages and distribution over different machines.

TCP and UDP-Sockets for Streaming

Alternatively to native integration, there is communication via network. This enables different instances of MobileSSI to run on individual machines while still operating synchronized and

sharing an overall pipeline. TCP or UDP sockets are available in MobileSSI to send continuous streams to other instances or application parts. These network streams have a predefined data type, such as floating point or integer, as well as a predefined dimensionality.

Therefore, sending and receiving heterogeneous data packages such as events is not supported via SSI.

OSC-Sockets for Events

OSC sockets have the benefit of defining the data structure that they transmit in their header. This enables sending/ receiving events as well as streams. Events are important because, in the case of classification results, they convey the actual interpretation of the input data, which is most useful for higher-level parts of an application.

Websockets and Web-Interfaces

An additional network socket to communicate with web applications is a websocket. MobileSSI's websocket implementation is based on Mongoose [3] and also provides a web server. The results

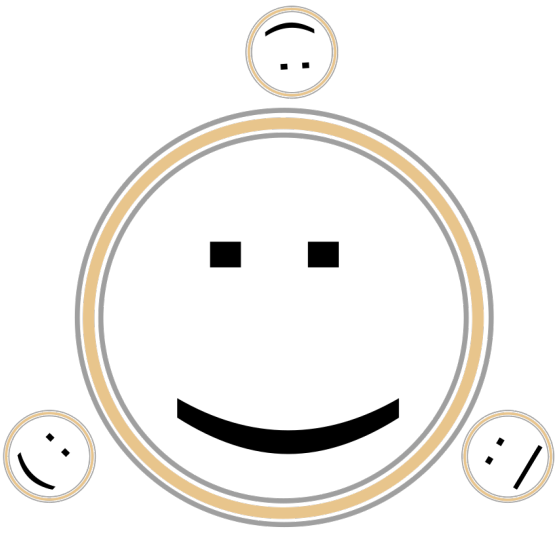


Figure 4.10: Emotion visualization UI connected to MobileSSI backend

of this allows the hosting of interfaces, of which the simplest form is provided by static pages' URLs via the stimuli plugin. Those interfaces are not limited to experimental settings but can provide feedback to the user or whole applications. As a result multi-modal fusion can be visualized to a group of users via a web interface see Figure 4.10.

[3] https://github.com/cesanta/mongoose

Here an interplay of different MobileSSI instances on different smart devices using Web as well as OSC sockets is needed. To find the setup described in detail, see Section 5.3 and Figure 5.7.

An example of a whole application can be found in a headache diary in Figure 4.11. It is not hosted by MobileSSI's websocket plugin, but deployed as a Tizen WebApp on a Samsung Gear S2. Nonetheless, the communication is realized via websockets and allows for complex label-acquisition, synchronized with data collection. Besides the origin of the pain in the first two

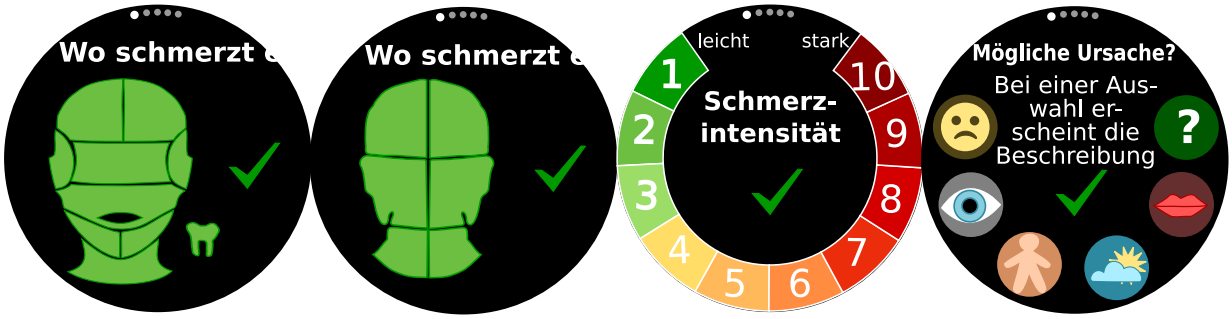


Figure 4.11: Headache Diary UI connected to MobileSSI backend

screens on the left, the intensity is recorded on the third screen and the possible cause on the fourth screen. Websockets allow the integration of MobileSSI into contemporary Apps that can be created rapidly for a variety of form factors.

Physical Computing

Instead of using touch input and graphical user interfaces, activities of daily living can be augmented using physical computing.

Figure 4.12: Smart Objects for Annotation "in the wild"

To ease annotation ”in the wild”, behavior can be tracked through the objects we interact with. For example, motion sensors on one pen can provide annotations for smart-watch-based learning of writing or drawing that could later be transferred to other pens without motion sensors. Communication between smart objects and MobileSSI is realized through Bluetooth Low Energy (BLE). For the tracking of drink behavior, a smart scale is used as drip mat, logging the event of rising a glass.

4.9 Summary

MobileSSI provides a port of SSI to UNIX platforms and Android, as well as additions for MSSP that handle new challenges in annotation and machine learning ”in the wild”. Meanwhile, MobileSSI is integrated into the code base of SSI. It expands the capabilities of SSI in rapid prototyping, using XML pipelines by HTML5 GUIs and carries over approaches to multi-modal fusion from the desktop PC to mobile devices.

The implementation of SSI is basis of experiments described in the following chapters. At the beginning in Chapter 5 a mobile implementation of multi-modal enjoyment recognition is realized, in Chapter 6 interactive machine learning for the recognition of drinking activities is introduced and finally in Chapter 7 the influence of local climate zones on well-being is examined.

Chapter 5.
Multi-Modal Laughter Recognition

MobileSSI has the goal of utilizing Mobile Social Signal Processing for the recognition of well-being in the wild. Laughter forms a paralinguistic social cue that is an objective to research in Social Signal Processing. Furthermore, it is an important marker for emotional and social wellbeing and thus will be described in this chapter.

To test MobileSSI's fusion capabilities within a familiar scenario, that has been subject to research on multi-modal emotion recognition [109], multi-modal laughter recognition forms the first endeavor of this thesis. This Chapter is based on the publications *MobileSSI - A Multi-modal Framework for Social Signal Interpretation on Mobile Devices* [65], *MobileSSI: asynchronous fusion for social signal interpretation in the wild* [66] and *Laughter detection in the wild: demonstrating a tool for mobile social signal processing and visualization* [67]. The own contribution lies in providing support for mobile platforms to the software framework, recording and labeling data in

Figure 5.1: Mobile setup: Three smart phones placed in breast pockets, clip-microphones.

the wild as well as evaluation using machine learning. Moreover, the employed fusion approach was extended to multiple mobile devices and users and presented with the implementation of a demonstrator.

5.1 Conception of multi-modal, mobile laughter recognition

While the setup in the lab involved video and audio as modalities and rarely sensors beyond [95], the mobile setup had to rely on a different set of sensors. While the audio input is still available, video recording is too intrusive and artificial to be considered for group enjoyment recognition with mobile devices. Since accelerometers are built into smart phones and are widely used e.g. in the human activity recognition community, they are a natural choice. Since body movement is important in laughter recognition [115] (Body Laughter Index), having the smart phone record chest-movement from within a breast pocket seems plausible. This is also supported by Consentino et al [39] that mention auditive, facial and body movement cues as characteristic of laughter behavior. They also attest, that most recognition systems rely only on audio, e.g. the work by Hagerer et al. [85, 86]. Fusion approaches rely mostly on audio visual information to date. A custom build device by Lacio et al. [48] forms an exception, where movement and physiological signals are fused using a wrist band. The work presented in this chapter, to the best of my knowledge, provides the only solution, fusing chest movement and auditive cues on an mobile device.

SSI's proven fusion algorithms are tested for their applicability in the wild, running on mobile devices that do feature extraction and classification in real-time, while employing asynchronous fusion over wireless network.

To sum this chapter's efforts, a demonstration with visualization is created, that uses HTML5 to present the recognized group enjoyment state to the user.

Thus, this chapter has four goals listed below:

- Switch to modalities common-place in MSSP

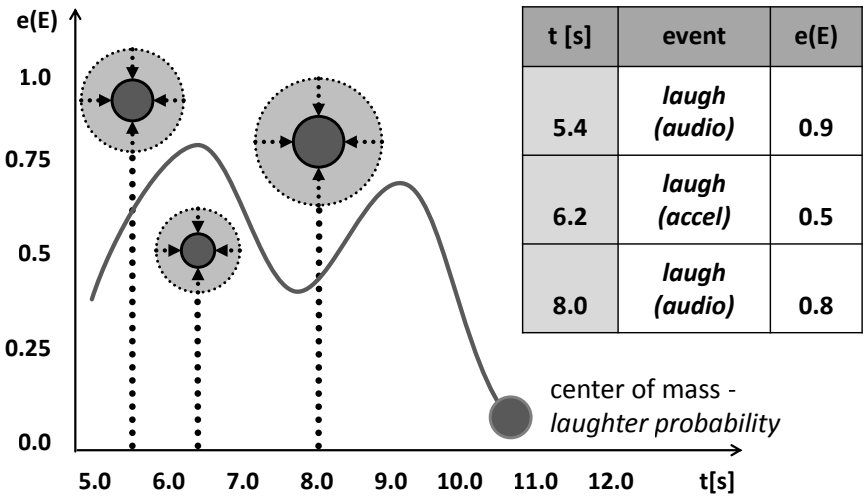
- Employ SSI's advanced fusion algorithms
- Create a realtime recognition pipeline for multiple users
- Implement a live demonstration with visualization

5.1.1 Fusion techniques

Fusion of multiple signals and modalities is key challenge of MSSP. There is a range of solutions to choose from, that apply at different stages of the processing pipeline and therefore bring different restrictions and possibilities.

Feature based fusion happens early in the process and enables to create one model that classfies across several modalities, whose features are considered at once.

Decision based fusion happens later, on the predictions of individual models trained on the according modalities via the features computed on their raw signals. There is an increase in flexibility, what set of rules is used to combine those results. Nonetheless, the modalities have to follow the same cycle in the sense of frame size. The models have to deliver classification results in sync, delay from wireless networks hinders consideration of multiple users.



t [s]	event	e(E)
5.4	*laugh (audio)*	0.9
6.2	*laugh (accel)*	0.5
8.0	*laugh (audio)*	0.8

Figure 5.2: The fusion algorithm considers the temporal flow of occurring, laughter indicating events. Influence of events decreases over time and a continuous laughter probability is calculated as the moving center of mass of weighted events. *Courtesy: Florian Lingenfelser*

Asynchronous fusion considers events over a longer time with individual weights and a continuous decay of events' influence on the fusion-process over time. This enables to not only consider modalities, that are evaluated at different pace, but take events into account that are contributed from other smart devices and their users.

Since human to human communication without technical support takes place in pairs or groups, it is close at hand to not just fuse one single person's modalities. The approaches to multi-person

fusion can be categorized into two classes. A more detailed look at group behavior uncovers, that group behavior is not just the aggregation of individuals' actions, but rather people that engage into a common activity and adjust to each other. Here synchrony is a reseached measure [214], synchrony is detected on signal or feature level, the implementation by Varni et al is not fit for real-time processing and as such of analytic nature. On a more abstract level, multi-person fusion can aggregate affective cues [40] within the Pleasure Arousal Dominance (PAD) domain of an model of emotion. Since it is not specialized on a single cue and the PAD-model is designed for a single person, the provided information is not well-founded.

5.2 Validation in the Wild

As a real world test, a laughter recognition study in an everyday setting is conducted. Laughter detection is a classic problem when it comes to social cues and SSI has already been used within an enjoyment recognition system based on audiovisual laugh and smile detection [110]. Data acquisition, however, was done in a typical stationary lab setting in which up to four study participants were recorded while telling each other funny stories of their lives [126]. Now, our aim is to port the existing system to run on mobile phones to investigate the following questions:

- Can the sensors of the previous system be adapted using solely sensor technology provided by mobile phones?
- Which parts of the SSP pipeline need adaption to work in a less predictable and changing environment?
- Can a comparable recognition performance be expected?

In principle, cameras could be used again for detecting visual laughter. However, they would have to either be place in the environment (which would limit the user's mobility) or have to be attached to the user. In the latter case, only visual laughter of the user's interlocutors could be captured. Of course, the camera could also be attached in a way that it faces the user. However, this setup would result into a rather bulky device. Consequently, another solution had to be found. Accelerometers seem to be a promising option. Indeed there is evidence from previous work using visual markers [115] or a complete motion capture suit [125] that motion is a good indicator of laughter. Therefore it was decided to replace the Kinect cameras that were used in the previous lab setting with accelerometers. For the audio modality no replacement was necessary since external microphones can be used to circumvent interferences from the pockets. The advantage of the new setup is that it uses only hardware that is available on smart phones or very easy to attach (microphones) and can be continuously assessed and analysed.

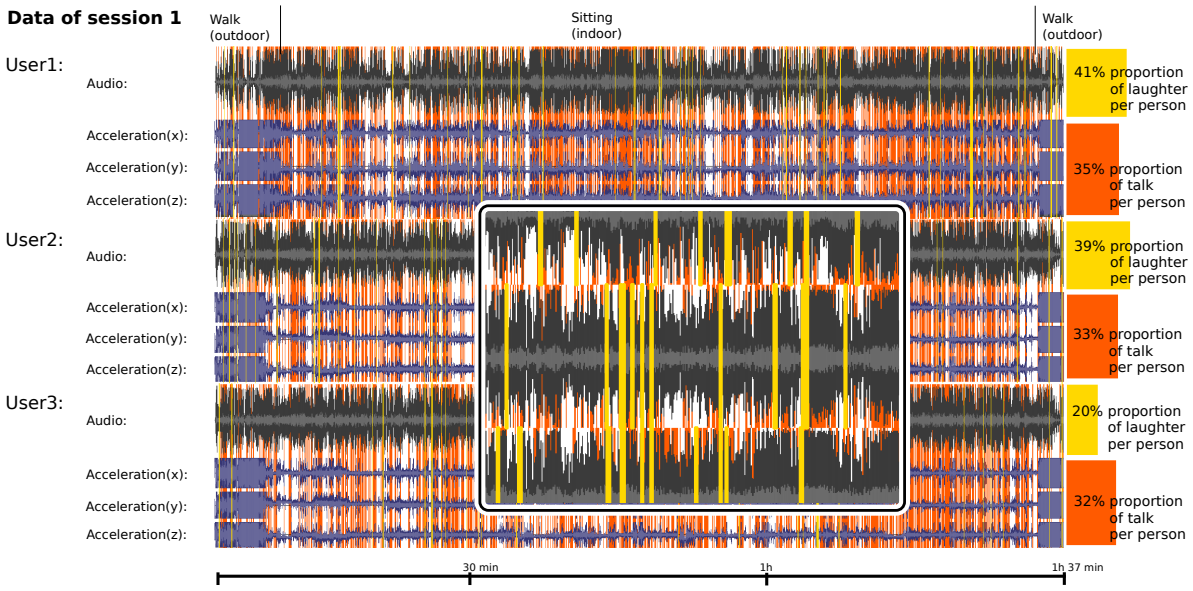


Figure 5.3: Overview of one session. Raw data containing audio and acceleration are plotted synchronized. Laugh (yellow) and talk (orange) events are marked and their proportion per user can be found on the right. The detailed window shows synchronised laughter between users.

5.2.1 Corpus

As a natural environment for the study a pub was chosen, as it is a common place for people to meet and have enjoyable conversations. As described above, audio and accelerometer sensors are used. The new setup is depicted in Figure 7.1 and shows three study participants, each of them equipped with a smart phone in his breast pocket connected to a clip microphone. The participants were acquired beforehand and given a brief introduction on how the setup worked. Apart from starting the session, no further interaction with the system was required from the participants. Throughout the session the participants were completely free in choosing the topic of their conversation, i.e. there were no guidelines were given on the content to be discussed.

Audio was recorded at 16 kHz, as it is the sample rate delivering the most reliable results on our target system vs. 48 kHz in the reference study. Accelerometer data were sampled at 100 Hz. Figure 5.4 features a signal snippet showing speech followed by a laugh event. For the study Samsung Galaxy S4 (GT-I9505) phones running Android 5.0.1 (latest official version at that time) were employed.

First, a pipeline was set up to continuously record audio and accelerometer data and relied on SSI's synchronisation techniques to ensure that captured signals are kept in sync (see [222]). Two sessions on different days were recorded and a total of four hours of natural conversations per user was collected. The experiments showed that data can be reliably captured with the sensors provided by the smart phones for up to eight hours per charge. Feature processing of six hours and online recognition for seven hours is possible with one charge. When reviewing the data shown in Figure 5.3 a significant amount of laughs was found (about 50 events per session

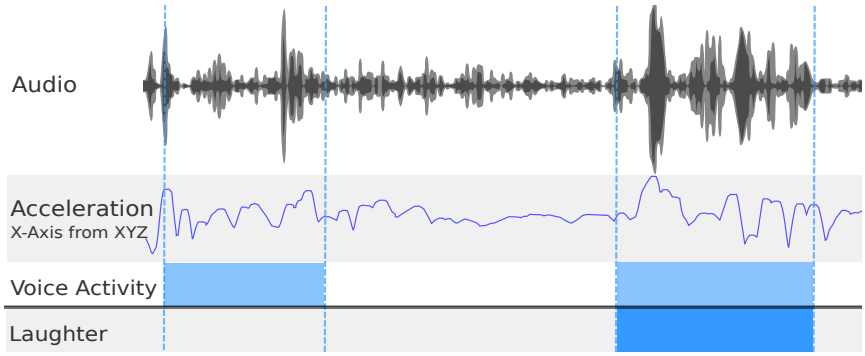


Figure 5.4: In addition to audio analysis acceleration is captured, which is an indicator of body movement to differentiate laughs from talk.

and user). In total 21500 overall samples were extracted by using a sliding window of one second and 400 ms frame shift whereof 875 contain laughter. In comparison, the corpus acquired in the reference study [110] contains 27000 samples with the same window and shift. Audio was used as ground truth to annotate laughter on both modalities. Figure 5.3 also shows that laughs are indeed infectious. Laughs of one person (indicated by the yellow lines) is immediately followed by the other two interlocutors.

5.2.2 Features

In order to recognize cues for laughter in the observed channels, relevant features had to be extracted from the segments of raw data. For audio, the EmoVoice feature set was used (1451 in total) [219] - containing MFCC, pitch, energy and more. For laughter recognition in audio data MFCC have proven themselves - not surprisingly as they are a useful tool in speech recognition and laughter has a lot in common with phonemes. For accelerometer data, a series of nine features was employed (listed in the enumeration below) for each of the three axis. The first and the second derivation for each calculated feature was added, resulting in a feature vector of size 81 for the accelerometer modality.

Features used on the raw signal, per axis, as well as their first and second derivation:

- Mean, Standard deviation
- Minimum, Maximum, Range
- Zero crossing rate
- Peak count, Pulse rate
- Energy

5.2.3 Evaluation

Evaluation is carried out frame-wise over the two recorded sessions. As both sessions feature the same users, it was decided to use two persons for training of recognition/fusion systems and keep the third for testing, thus having a fixed training and test set. This evaluation approach *leaving one user out* simulates the performance of an online system and allows us to draw a direct comparison to the reference system [110] evaluated the same way. To get a first impression of recognition performance, one SVM-model for each of the two modalities is trained, audio and smart phone acceleration separately with two classes (laughter and talk). Frame-wise recognition results are shown in Table 5.1. The tables present unweighted recognition results (average accuracy across classes), because the number of frames actually containing laughter is of course considerably lower than frames that show no hints of laughs. This prevents high detection rates by only favoring the dominant class (weighting the average with the classes sample count).

	Uni–Modal Classification	
	Accelerometer	**Audio**
Talk	63.42%	86.70%
Laughter	80.95%	76.19%
Average	**72.19%**	**81.45%**

Table 5.1: Results of classification per modality.

While the reference system reached an unweighted accuracy of up to 90 % for laughter recognition on audio frames, now a clear drop to 81 % in recognition accuracy can be observed. The detection rate for the accelerometer data was lower, too, yielding 72 % compared to 79 % obtained with the video modality in the laboratory study.

In order to compare the performance of the proposed event-based fusion approach a very basic decision-level fusion strategy is applied also (Table 5.2). Decision-level fusion using the product rule [221] improved the results by one percent point over uni-modal classification and scored 82.59 %. As a second method, asynchronous fusion on event-level features (Section 5.1.1) was conducted and improved the classification by three percent points to 84.64 %. Instead of fusing information over fixed time segments, it is decided frame by frame how much recognized events should contribute to the fusion result. To this end, the following parameters are taken into account. Each event is assigned a modality-specific weight to emphasize more reliable information sources. A decay parameter determines how fast the influence of events on the fusion result decreases (see Figure 5.2). If a particular threshold is achieved for a particular frame, the frame is classified as laughter. The optimal configuration of these parameters was learned on the training data by systematically testing a large number of parameter combinations following the grid search approach described in [110]. Within 12000 combinations of parameters, based on our previous research and additional adjustments for the new setting, 18 configurations were

	Multi–Modal Classification	
	Decision Fusion	**Event Fusion**
Talk	86.62%	85.95%
Laughter	78.57%	83.33%
Average	**82.59%**	**84.64%**

Table 5.2: Results of classification fused using decision- and event-driven solutions

found that scored an average of 84 % detection rate. These configurations give events from the audio modality a higher influence (0.7 or 1.0) while accelerometer events are weighted lighter at 0.1 to 0.3. Audio and acceleration decay parameters are comparable and vary from 0.6 to 1.0 (audio) and 0.5 to 1.5 (acceleration). This is plausible as audio is the modality with better classification results in Table 5.1, therefore can be relied upon more and faster, while accelerometer events make a better contribution if they are weighted less.

5.3 Demonstration within a Multi-User Scenario

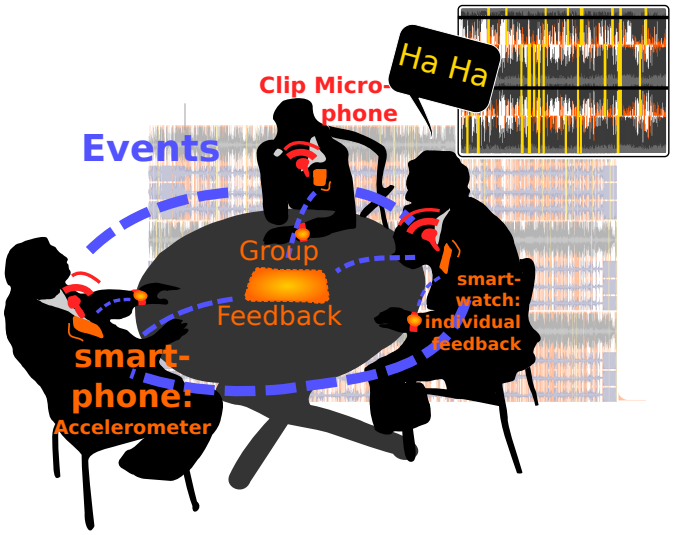


Figure 5.5: Devices involved in the group enjoyment recognition with multi-user feedback

There are recent advances for uni-modal and single user laughter recognition tailored for mobile device usage [85], that still operate reliably with 50% of droipped frame [86]. Multimodal solutions with asynchronous fusion, as the one we presented, can still operate correctly if one modality is fallen out.

MobileSSI allows to build applications that fuse not only multiple modalities but also streams captured of multiple users. The demonstrator features inter-personal fusion to obtain a group enjoyment level.

Figure 5.6: Devices used for information visualization in the laughter demonstrator

Figure 6.8 shows the components of a pipeline that has been created for the laughter demonstrator. It includes mobile audio and accelerometer sensors, transformers applying features on the captured signals as well as consumers for classification and output. To orchestrate a multitude of mobile devices, the following components have been added to the pipeline:

- *SocketEventWriter*: sends events over network
- *SocketEventReader*: receives events to fuse them in the recognition pipeline
- *Websocket*: provides an HTTP and WebSocket server used to host a web page on the device that can be displayed on smart watches and the tablet

5.4 Demo Setup

With the laughter demonstrator, smart phones are combined as sensing devices , smart watches as personal and tablets as group displays. While users engage in social interactions with each other, MobileSSI records and synchronizes data from the audio and accelerometer sensors embedded in the mobile smart phones worn in the users' breast pocket. Using an event-driven fusion approach, the users' audio and accelerometer data are integrated to determine their degree of enjoyment. Smart watches and tablets display information on aggregated cues of enjoyment at the individual and the group level in real-time (see Figure 5.8).

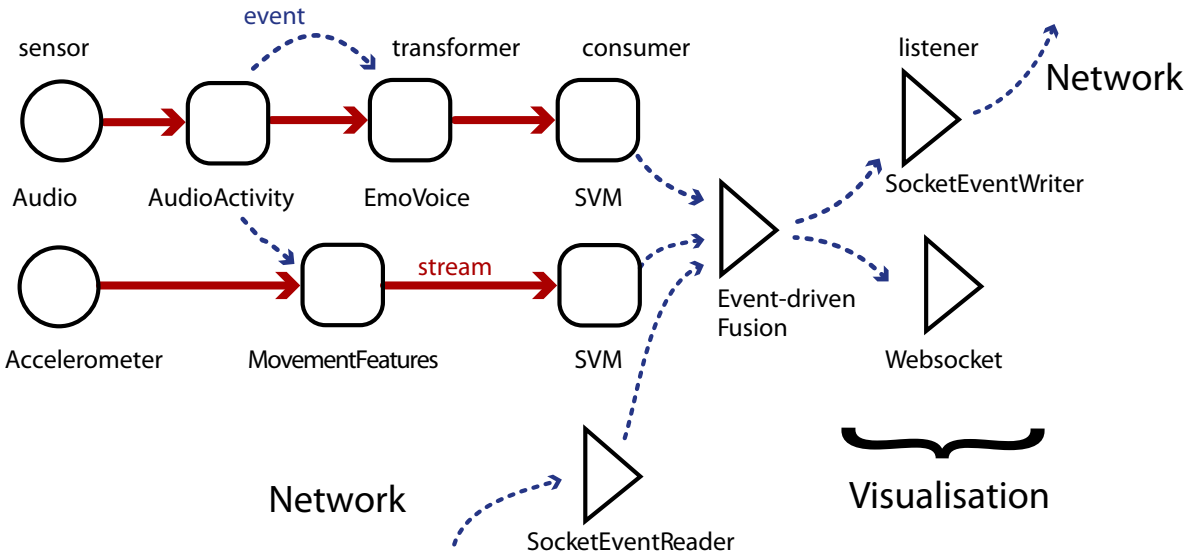


Figure 5.7: Sketch of the recognition pipeline (signal flow from left to right): features are extracted from the raw streams when voice activity is detected in the audio channel. Support Vector Machine (SVM) classifiers recognize the presence of laughter in the channels. Laughter events from both modalities are fused with events received over the network and visualized through the websocket interface.

Figure 5.8: Visualization of enjoyment at the individual (left) and the group level (right)

5.5 Discussion

Compared to the story–telling corpus there is a clear differences regarding the signal quality. For instance, in the pub the captured audio signals were overlaid with diverse sources of noise: music playing in the background, surrounding conversations of varying intensity, utterances of the waitress while taking orders, interferences with mobile network activity etc. These disturbances present great challenges to voice activity detection and audio classification and should be addressed, for instance, by applying noise reduction techniques. Since the environment in a mobile setting is subject to great changes, e. g. when the group is temporarily leaving the pub for a smoke (see Figure 5.9), noise cancellation schemes are required that are able to dynamically adapt to the current situation. It is important to note that the surrounding sound scape also contains relevant data that should be analysed to gain further information about the envi-

ronment and the user's activity. For instance, tailored classification models could be used for outdoor and indoor settings.

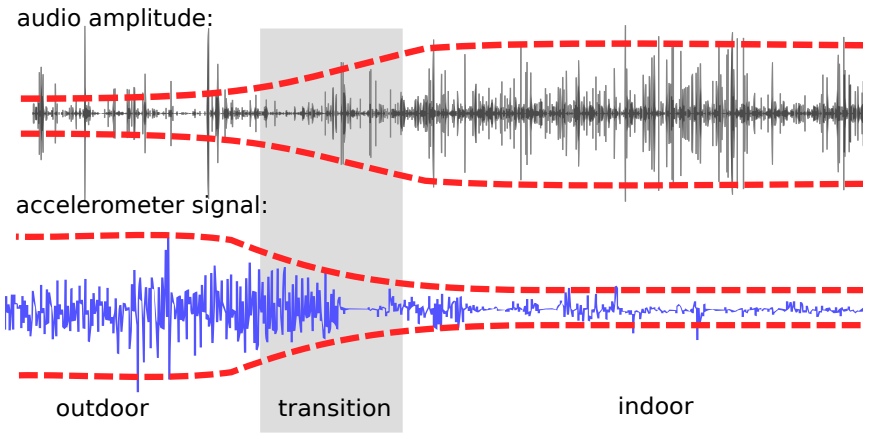


Figure 5.9: Change in audio amplitude and accelerometer energy before and after entering the pub.

Overall, the experiment demonstrated the benefits of MobileSSI when porting existing lab settings into a mobile environment. Presented classification results are clearly lower than those obtained in the lab. However, techniques based on event fusion narrow the gap compared to uni-modal classification. The smile and laugh detector by Fukumoto et al. [76] obtained recognition rates of 89.2 %. However, they had study participants watch videos of ten minutes only while we investigated social interactions over several hours in a mobile setting. Also they used a setup with glasses equipped with photo interrupters and relied on a PC for online processing whereas in the presented case all the computing is done on the phone. Since battery life of today's smart phones is sufficient to record and process data in real-time for several hours, MobileSSI is able to run real-life experiments, which provide better insights on the actual challenges that have to be faced when applying social signal processing "in the wild".

5.6 Summary

The mobile adoption of SSI proved to be successful in using acceleration as modality in laughter recognition. A setup evaluated by Lingenfelser et al. [110] in the lab relying on video and audio as modalities for group enjoyment recognition served as starting point. Since video can not be used as modality in the wild due to intrusive setups, it was replaced with the accelerometers of smart phones worn in the users' breast pocket. Chest movement provides a source of information next to audio, that contributes considerably to a joined recognition process. Functionals, a collection of statistical features were used in the machine learning process and SVMs served as classifiers. Acceleration is generally the weaker modality in laughter recognition, scoring almost 10 percent-points lower that audio, but reaching higher accuracy on the laughter-class compared to audio, with 80.95 % to 76.19 %. This enabled late fusion such as the application of

the product rule, but especially the event-based approach to score higher than the individual modalities at 84.64 % compared to 81.45 %. Furthermore, it was possible to adapt the asynchronous fusion strategy to fuse information of up to three persons to estimate the level of the group's amusement. In a live demonstrator the visualization of group enjoyment could be achieved on basis of online feature-extraction and classification. This was realized by distributing multi-modal, event-based fusion over multiple devices using OSC-Sockets and using a self-hosted web-application, connected via websockets.

Chapter 6. Interactive Training of Drink Activity Recognition

In M-health, nutrition is the field of use with second most apps (after fitness apps) under teens and young adults in the US [166]. Yet, nutrition apps make the user log a lot of activity manually [185]. To support the user in the activity of nutrition logging, the fluid intake could be logged with drink-activity recognition. Since drinking is an activity that can vary in execution depending on the context, habit and stature of the user, a personalized model is desirable. Here interactive machine learning (iML) comes into play.

6.1 Conception of interactive, mobile machine learning of drink-activity recognition

Interactive machine learning allows the adaption of an activity recognition system that is based on machine learning. Next to personalization, it has the advantage of training on the user's device, to increase privacy. Interactive machine learning forms an addition to MobileSSI and can be used in different scenarios.

Drinking as behavior, forms an entry-point for augmentation. This allows to hook applications, such as health apps, into natural behavior and thus create a seamless integration of computer interaction. Furthermore, it allows health application to interact with that behavior to support the user in making a behavior change and therefore improving his health.

This chapter is based on the publication *DrinkWatch: A Mobile Wellbeing Application Based on Interactive and Cooperative Machine Learning* [68], where own contributions lie in the implementation of interactive machine learning capabilities into MobileSSI, creating a corpus in the wild, with annotation, employ machine learning and simulation for evaluation as well as conducting bodystorming with test users.

6.2 Background

Arguably, the advent of mobile and ubiquitous technology has disrupted how we (as users) envision technology's role in our everyday life. While originally mobile devices were perceived as personal information management tools, and thus as *tools* in a traditional sense, today's mobiles have access to a vast amount of knowledge from which they can learn, and seemingly become a companion, we become increasingly depended on.

There are some benefits of this ongoing shift of agency and capabilities towards mobiles or technology in general, such as technology becoming able to recognize harmful behavioral habits of users and assist users in reflecting on their habits and hopefully provide support in adopting positive habits. Be it to regularly taking a walk or drinking enough, behavior change bears great potential towards improving wellbeing.

In the following, related work in human activity recognition is summarized, which is an essential part in recognizing human behavior, and describe different ML approaches with regard to their characteristics and application domains.

6.2.1 Human Activity Recognition

Over the last two decades, research in Human Activity Recognition (HAR) has been focusing on a wide range of applications, such as surveillance and security [201], ambient intelligence [168]

(e.g. to assist older adults [207]), or health care [231]. In particular, in ubiquitous computing environments or smart home environments, Human Activity Recognition is a key feature, for example, to monitor daily activities of users or provide assistance [237].

The rapid technical development of mobile devices and wearables, such as smart phones and smart watches, has further expanded the possibilities for HAR. Mobile devices are equipped with a plethora of sensors and are worn or carried around all day. Thus, many activities of users can potentially be recognized. Consequently, a lot of research that investigated methods and applications [230] for HAR has emerged, in particular, research employing inertial sensors of smart phones [131, 197].

In more detail, HAR is used to automatically recognize a person's activities from a stream of sensor data, for example to pro-actively provide assistance, log daily routines, or to initiate necessary procedures (such as calling an ambulance or neighbors in case a person has fallen [23]). This makes them an important entity among today's E-health topics, be it detecting stereotyped movements in children with developmental disabilities [107] or automatic monitoring of rehabilitation processes [203] or using smart cups to track the behavior of residents of an inpatient nursing care facility [239]. In comparison to smart phones, smart watches have a decisive advantage, which makes them particularly suitable for HAR. They are body-mounted and therefore always at the same place (i.e. constantly attached to the user's arm throughout a day). The human arm is actively involved in most of daily activities, whereas movements of the body can be smaller and may only reveal few activities.

Smart phones usually detect only movements related to the whole body due to their typical placement in the pocket. Therefore, the number of identifiable activities with smart phones is limited. Examples from the literature include walking, running, jogging, standing, sitting, walking up/down stairs, or using an elevator [79, 131, 134]. In contrast, smart watches or wrist worn wearables have the potential to detect more activities than with smart phones, such as drinking, smoking, typing on the keyboard, or eating with a knife and fork. Thus, new application areas can be addressed, such as food/drink reminders and related habit awareness applications (e.g. [163, 197]). The fact that smart watches record the subtleties of each individual's arm movements in turn allows ML algorithms to generate personalized models for activity recognition. Personalized models usually result in higher precision of recognition algorithms and require less amount of sample data than user-independent models.

With the rapid development of smart watch technology, HAR on smart watches is an ascending topic [22, 194]. In contrast to previous work that utilizes smart watches, the work at hand combines online learning and interactive ML to continuously improve activity recognition models. Moreover, the complete learning process is done solely on the smart watch without access to any online resources or requiring network connectivity.

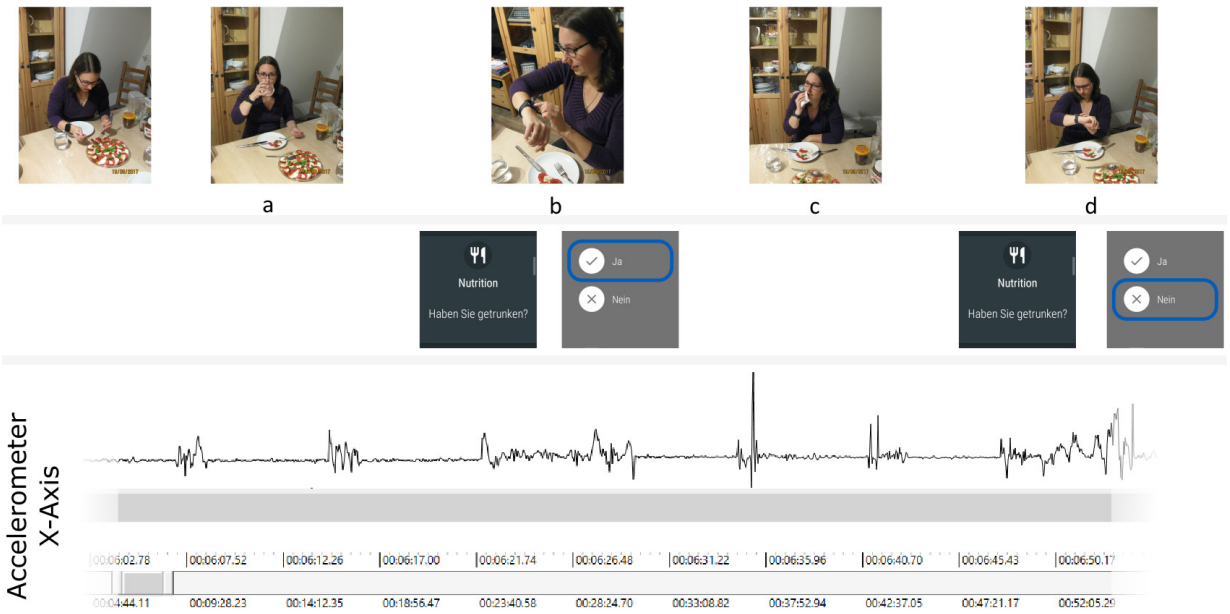


Figure 6.1: An exemplary health application scenario presenting the interaction and cooperation between a user and the DrinkWatch application. The second row provides screenshots of the DrinkWatch application and the third row presents raw accelerometer data of one movement axis as exemplary sensor data, which are used to recognize the drinking activity.

6.2.2 Active Learning

Miu et al. [131] presented an Online Active Learning framework and studied how to collect user-provided annotations to bootstrap personalized activity models. They demonstrated that generating personalized Human Activity Recognition models can be achieved on-the-fly and does not require expert supervision or retrospective annotation of sample data. While Miu et al. made use of a smart phone app to query the user, the work at hand queries annotations through a smart watch interface. A smart watch app has the benefit that queries on a smart watch can be handled more comfortably and quickly since smart watches don't require users to get it out of the pocket first.

Active Learning has been investigated for different models and classification types (e.g. Support Vector Machines [208]) as well as different types of query strategies. An overview is given by Settles [192]. The most widely used approach and therefore selected as entry point in Section 6.3.3.5, is Uncertainty Sampling [108] also called Query on Uncertainty. In these approaches, a labeling system picks up samples for which the target class cannot be determined with a high certainty. This way the system is not locked into just learning from data that it already handles well. According to Lewis et al. [108], this approach performs better than relevance sampling which picks high confidence samples for relabeling.

Also common is the approach called Query by Committee [193], where multiple models, that have a strongly different way of operating, are grouped into a committee. Those samples are

chosen for labeling by an oracle, where the individual models disagree most. Other query methods are focused on error reduction [192]. They either directly try to maximize the expected error reduction with the selected sample or they look at the expected model change that is expected from all possible labels of the sample.

6.2.3 Interactive Machine Learning (iML)

Fails et al. [64] iML within perceptual user interfaces to allow the user to train view and correct classifications. They discussed their approach within the scenario of image processing and object detection. Interactive Machine Learning gives the user control over systems that else are intransparent and inadaptive [5]. While it is not widely adopted, Gilles et al. see an important role with iML in movement design [80]. This is connected to the property of accelerometer data, that are much harder to interpret by themselves without audio-visual reference. Nonetheless, motion data are important in mobile computing e.g. within the context of drink activity recognition as proposed in this chapter.

Few recent works on HAR have investigated iML based on a smart watch [194, 197] or online learning with a smart watch [131], but no one has attempted to combine both approaches to realize interactive online machine learning solely on a smart watch independent of external computing resources.

Shahmohammadi et al. [195] use a smart watch for interactive machine learning, since smart watches can be worn and allow the display of user interfaces. Active learning is used to identify five common activities of daily living. The data are not processed on device but send to a web server. The activities (Walking, Standing, Sitting, Lying Down, Running) are not selected to serve within the context of a certain task or application but to prove the concept. Study participants were asked via smart watch interface to perform certain activity instead of capturing natural behavior that is labeled by the user. Active learning was not run asking the user directly but used the collected data for label requests.

With DrinkWatch, design and implementation of an application prototype for smart watches is presented, that combines both, local processing and active learning. Furthermore, insights from a technical evaluation are shared.

6.3 DrinkWatch Prototype

Across all the state-of-the-art and off-the-shelf mobile devices, smart watches seem most suitable in providing least intrusive and immediate feedback in mobile settings, and thus, allowing users to reflect on their immediate activities and contextual habits. While their form factor and small size is indeed an advantage when considering their integration in everyday situations, it

is also often challenging to design and to develop interactive applications for smart watches. For example, smart watches provide only a very small-sized screen which limits the amount of information that can be presented to users. This limitation is, however, not relevant for the intended use case of DrinkWatch as it mainly makes use of the movements of the smart watch for hand activity logging. The presented system only occasionally shows notifications to users and asks them for feedback related to activities. The prototype system further aims to reduce the complexity and amount of interaction (required to recognize and log drink activities) through automation.

DrinkWatch aims at recognizing drink activities (by means of inertial sensor data of the smart watch) and tracks each drink activity for later analysis (see Figure 6.1). If DrinkWatch senses "interesting data", which potentially represent a drink activity worth learning from (Figure 6.1a), the smart watch queries the user for assigning a label to the recorded sample data (Figure 6.1b). Thereby, the user is actively involved in the ML process and may choose to adapt the drink activity model or not. Consequently, not only drinking, but also activities, such as blowing one's nose or wiping one's mouth (Figure 6.1c) may lead to a query to the user (Figure 6.1d).

DrinkWatch serves three main functions.

- First, it offers a graphical *user interface* for querying the user for annotations and for reviewing recognized/logged activities.
- Second, DrinkWatch continuously collects data samples from the watch's accelerometer and other potential data sources. In the prototype, a smart scale is included, as outlined in Section 6.3.2.2. This data collection, the *corpus* (Section 6.3.2), serves as the basis for a *warmstart model* in the ongoing classic machine learning (cML) process. For the purpose of later evaluations, all collected data samples are also locally logged on the smart watch. However, this is not required for the online learning approach since the learning process requires only the latest annotated sample, see Section 6.3.3.4.
- Third, Drinkwatch integrates an *ML logic*, which runs as a service on the smart watch. While most of the logic, such as the online learning classifier, are implemented in the C++ programming language, part of the logic is embedded in a thin Java layer connecting the ML logic with the Android system (e.g. user interface) via JNI.

In the following,each of the three parts of the DrinkWatch are described, including the implementation of the ML logic (see Section 6.3.3) in detail.

6.3.1 User Interface

DrinkWatch is implemented as a stand-alone application that runs on the smart watch Asus ZenWatch 2, which is using the mobile operating system Android Wear 2 (Figure 6.2). Beneath

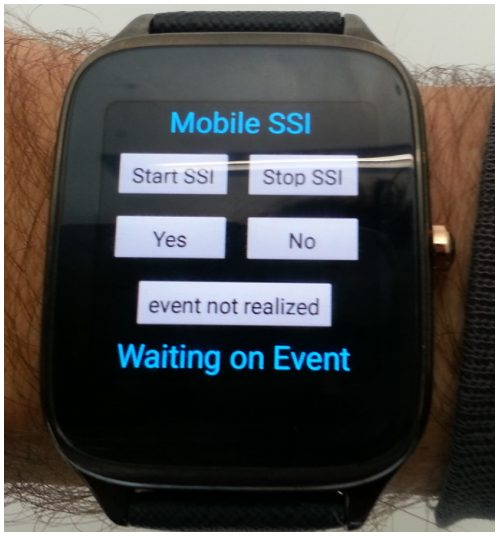


Figure 6.2: Cooperative Learning Interface on the smart watch. The first two buttons enable the user to start or stop the recognition pipeline. Whenever a drinking activity is detected, the user can inform the system whether the recognition was correct (”Yes”) or incorrect (”No”). Additionally, with the last button, the user is able to indicate whether a drinking activity was not detected.

the up-to-date OS, it can be charged and programmed fast using a USB connection, which is handy for development and experiments. There are hardware solutions with a wider range of sensors or fitted input hardware, such as a bezel, that might be more attractive for long-term use. A minimal user interface on the watch (see Figure 6.2) is used to handle queries to the user and to start and stop the learning pipeline. Thus, users have control over when and whether to provide labels. The simple interface allows non-expert users to easily provide feedback on the go. Drink activities that lie within the desired confidence range of the iML model trigger a request/notification. Notifications are given by playing the standard notification sound of the watch and displaying a text (”Have you been drinking?” instead of ”Waiting on Event”). In the current prototype implementation, the vibration function of the watch had to be turned off, since it influenced its accelerometer sensor. This issue will be solved in a next iteration by disabling sensor reading while a vibration is being executed by the watch.

6.3.2 Corpus for the Warmstart Model

6.3.2.1 Recording setup

In contrast to many other studies on activity recognition, people are not asked to perform specific actions, but rather sample data in everyday situations are recorded to be labeled afterwards based on a ground truth. The recording setup was slightly different from session to session. Recording of acceleration data from the wrist was always performed using an Asus Zenwatch 2. In addition, the setup also included a camera to record video of the user when

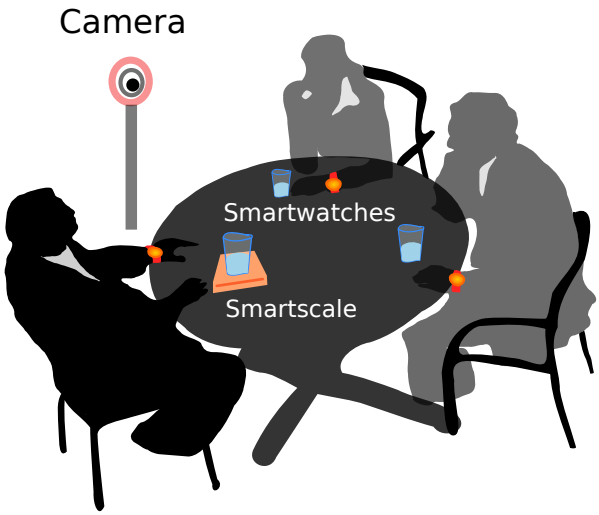


Figure 6.3: Recording setup with up to three people wearing smart watches to record labeled accelerometer data for the initial classification model. The weight of one person's drinking vessel acquired by a smart scale and video data were additionally recorded to be able to annotate drink activities afterwards.

possible. The number of users per session varied from three to one, while 22 sessions (out of 25) had only a single user (see Figure 7.1). All users were asked to wear the watch on their preferred hand. All recordings, except for five sessions, contained smart scale data, which can be used by our iML approach to speed up the annotation process.

6.3.2.2 Smartscale

The smart scale prototype [186] in the presented system (see Figure 6.4) continuously broadcasts weight data of vessels placed on it via Bluetooth 4.0 to every receiver that is nearby. In this case, the smart watch received and recorded the data whenever the watch was in reach of the smart scale.

Figure 6.5 exemplary shows recorded data of the smart scale. The graph resulted from drinking from two 0.5 l PET bottles (one by one). After each drinking activity the bottle was placed on the smart scale. When the first bottle was empty it was replaced by a full one. The plot shows that the first bottle was not completely full and has not been placed on the sensor, after being empty.

Whenever someone wants to drink out of a vessel placed on the scale, he or she usually first takes the vessel from the smart scale (weight is 0 g), drinks out of the vessel, and places the vessel back on the smart scale. The weight is now lower than before. The mass can increase if additional fluid is filled into a vessel or another vessel is being used which is heavier and/or contains more fluid.

In comparison to accelerometer data, the weight data of the smart scale is easier to interpret

Figure 6.4: A glass of apple juice standing on the smartscale.

so that an annotator can quickly detect a drink activity, but also enables automated annotations. The video data can be used to validate the labeled time segments but does not have to be completely watched.

6.3.2.3 Dataset

The data set contains 25 recorded sessions, which overall consist of 16 hours and 30 minutes of every day activities containing 5117 samples of drink activities and 26288 samples of non-drink activities. One sample consists of a 1 second frame step together with 7 seconds of overlapping preceding data. A typical snippet of a drink activity is shown in Figure 6.7. Such an activity is characterized by three phases: picking up, bringing the vessel to the mouth and back as well as finally putting the vessel down.

Random under-sampling was used to balance both classes in the training process. Acceleration data were recorded with 25 samples per second using the accelerometer sensor of an Asus ZenWatch 2. As ground truth video and smart scale data were recorded synchronously. An annotation session containing all data can be found in Figure 6.9. Furthermore, the Android system provides a so-called *linear acceleration sensor*, which represents the raw acceleration sensor exempt from the earth gravitation influence. The DrinkWatch prototype makes use of this linear acceleration sensor as it provides better performance for HAR [194]. These data were used to simulate a cML process and to gain a warmstart model for further iML, see Section 6.4. Thus, the data set is an important input for the ML module described in the following.

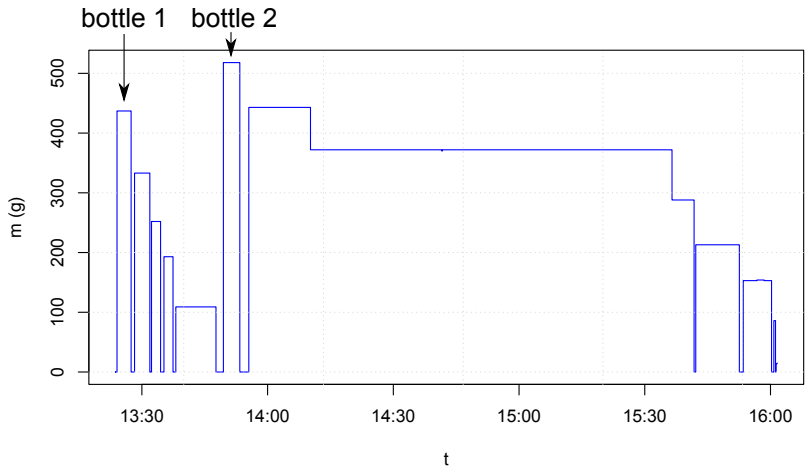


Figure 6.5: Weight data of the smart scale. Two filled 0.5 ml PET bottles have been drunken during this session. Whenever the drinking vessel is lifted the weight is 0 g (short lifting is omitted). After drinking the weight is reduced.

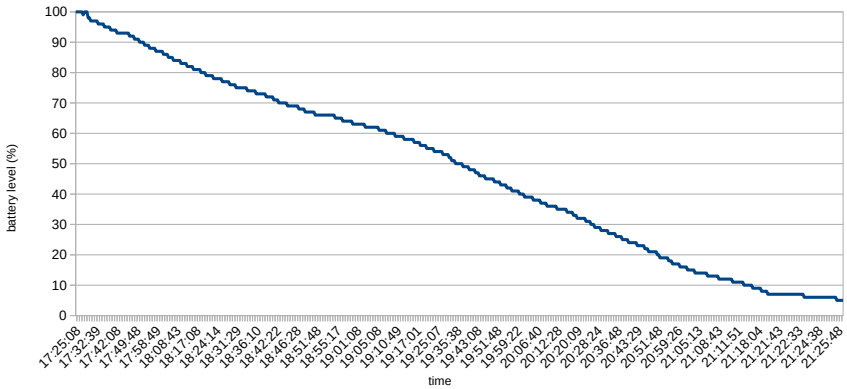


Figure 6.6: Battery level of Asus ZenWatch 2 running the MobileSSI iML pipeline

6.3.3 Implementation of the ML Module

DrinkWatch employs activity recognition to reduce manual logging effort that is required by the user when using a notebook or a conventional logging app. To this end, DrinkWatch continuously tracks the user's wrist activities in order to detect specific time windows (frames) that may be interpreted as an indicator of drinking. In case of high confidence, a drink event is automatically registered by a higher level app, e.g. a nutrition logging app. In case of low confidence, the system has to decide whether to ask the user for confirmation or not. Information gain is considered as well as the user's situation, as discussed by Amershi et al. [5]. For example, the user should not be disturbed if the expected information gain is very low.

The maximum run-time of the system without WiFi is about four hours, as can be seen in the graph of Figure 6.6. In case of low battery (2 %), the prototype app stops the ML pipeline in order to properly finish the session. From the two days maximum battery life under optimized circumstances, this means a strong reduction.

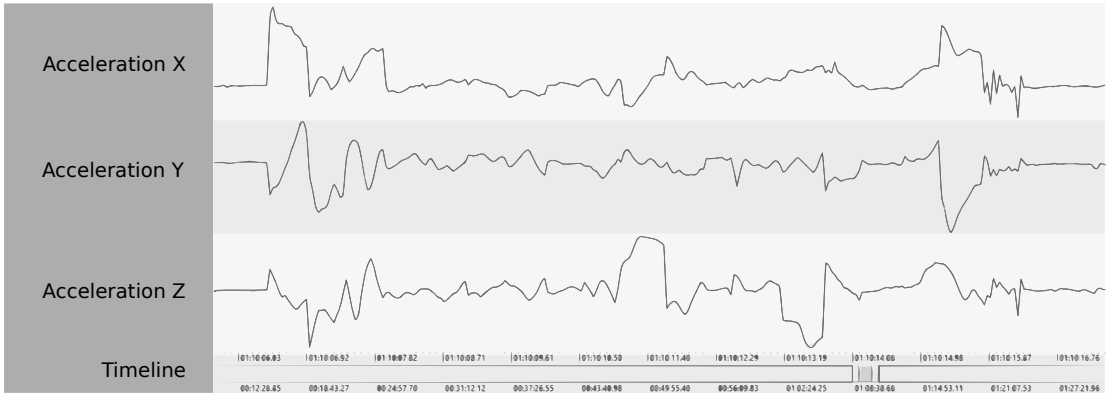


Figure 6.7: Three axis accelerometer data of a drink activity. The start and end of the signal describe the movement of the drinking vessel to and from the mouth. In the middle of the signal the rotation of the vessel by turning the wrist takes place.

The prototype relies on MobileSSI. . While MobileSSI already has ML capabilities for a range of classifiers, implementation followed a classic non-interactive approach. The extensions include online learning capabilities (see Section 6.3.3.4) that enable the user to interact with the model using a simple user interface while the model actively (see Section 6.3.3.5) queries the user. The prototype also shares parts with a classic ML pipeline, such as data collection and feature extraction, which are also described in the following. A brief overview of the pipeline and application concept is given in Figure 6.8. The red arrows mark continuous streams with a fixed sample rate kept in sync by the SSI framework. Blue dotted arrows mark events that are sporadic, but contain a time stamp and duration. Gray components are either future work, the user moderation and context component, or not described in this thesis, namely the integration with the nutrition logging app. User moderation, is an additional layer atop of activity recognition and requires an own model that relies on further sensors such as application meta data. First work into this direction was conducted by the author of this thesis with the nutrition app presented in Seiderer et al. [187], which allows the user to choose in each situation the most appropriate device combination out of a smartphone, smartwatch and smartscale.

6.3.3.1 Frame Size

In order to continuously process data, segmentation of the data has to be addressed. A fixed window is set to a size of 1 second together with an overlap of preceding 7 seconds. This allows the whole event to be captured, in most cases while having a reactive system, giving quick feedback. Given the chosen sample rate of 25 Hz 200 raw data points in three dimensions can be gained, as the accelerometer in use has three axes.

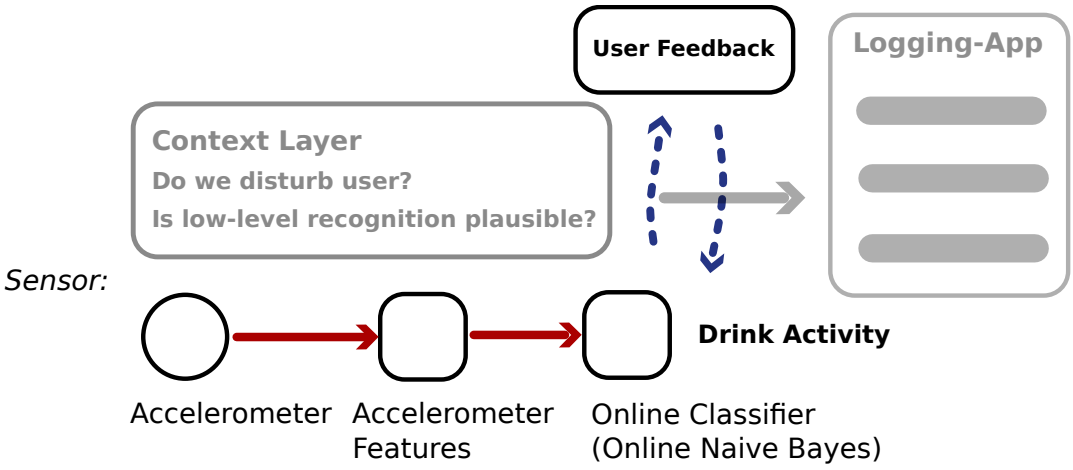


Figure 6.8: Overview over iML Pipeline and future system components.

6.3.3.2 Feature Selection

Accelerometer data are widely used in Human Activity Recognition and a lot of features have been experimented with. Features are needed to simplify the classification process since training classifiers directly on raw data typically uses more resources both in data demand and computing power. The selected feature set for DrinkWatch is based on related work. In particular, a range of features that are known to work well on acceleration data [13, 35, 93, 106, 161] and have been used for the recognition of drink activities are implemented [104, 226].

On each axis/dimension , the following features were calculated:

- Mean, Min, Max, Std. deviation, Variance, Energy
- Interquartile range (IQR)
- Mean absolute deviation (MAD)
- Root mean square (RMS)

Additionally following features are generalizing over all axes:

- Correlation between XY, XZ, YZ axis
- Mean, Std. deviation, Min, Max, IQR on length of per sample vector over all axes (magnitude)

This results in overall 35 features calculated on the previously described 1 + 7 seconds containing 200 samples.

6.3.3.3 Normalization

Normalization is scaling all features' data range to fit a certain range, in this case within 0 and 1. This is, for example, done by using the accelerometer's maximum output value that can be queried using the Android API. Compared to a classical approach in MobileSSI and many other implementations of classical ML, the responsibility of data normalization is moved from the training process, iterating over all samples in the data set, to the feature calculation, on the current chunk of data. This is necessary because with low initial sample count determining the minimum and maximum on already known data might not be representative for future data. There are alternatives to feature based normalization, such as adaptive scaling. While normalization is not strictly necessary for Naive Bayes based on a normal distribution, it is recommended to keep features with higher values from dominating features with small values. The pipeline provides a feature vector of dimension 35 that is fed into the following online classifier component every second.

6.3.3.4 Incremental Learner

Classification of the current data frame is handled by the pipeline, as it would be the case in a classic ML pipeline. The main objective is to continuously improve learned models for fluid intake based on tracked data and user input. Incremental [205] or online learning enables us to learn a new model from scratch in the deployed application. Furthermore, the model can be improved at the moment the user provides new labeled data and the next input can be analyzed with the improved model without the need to restart or stop the application. To speed up the process, a classically trained model is used as a starting point for further incremental training. This procedure is called warm-start.

Naive Bayes is the classifier of choice, which can be easily adapted for online learning (see e.g. the implementation used in MOA [24]). The online learning variant of Naive Bayes incrementally calculates mean, variance and standard derivation and additionally stores the sample count to be able to adjust with new data proportionally. The algorithm is described in detail by Knuth [102] on page 115. The calculations are executed per feature and class, thus the model consists of 210 float values and a sample count. As Naive Bayes classification results into confidence values, it enables us to query the user based on the level of uncertainty. Furthermore, it is fast in training and execution. This makes Naive Bayes a good option for restricted platforms, such as smart watches. Moreover, it offers an advantage in data security, as no other data that can give an insight in user behavior or health related information are permanently saved to the watch. At this point LibLinear is only integrated without online learning capabilities, but a solution exists according to Tsai et al. [210]. The future integration of LibLinear as additional online learning library depends on the result of the evaluation, see Section 6.4.

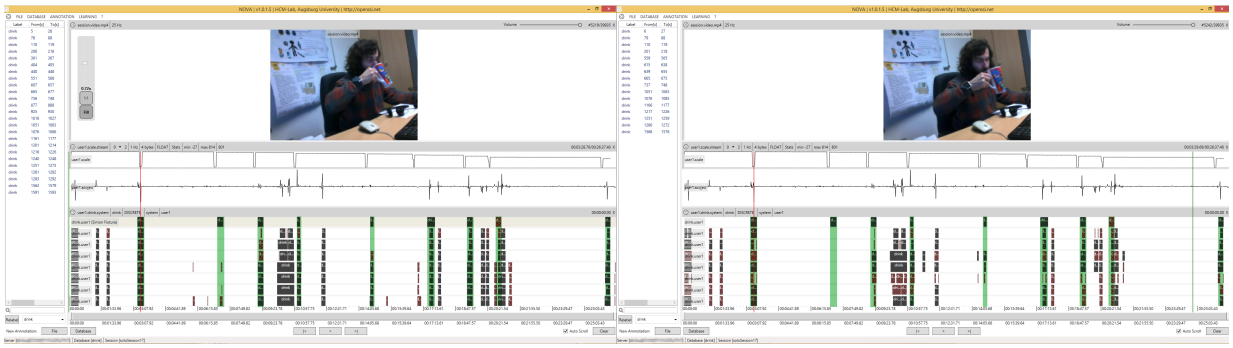


Figure 6.9: Cooperative Machine Leaning in NOVA: Predictions of LibLinear (left) and Naive Bayes (right) on one session. Video, smart scale and acceleration data are followed by annotations. The first line contains the hand labeled annotation and is followed by predictions of models with increased number of training data. Areas marked in green are drinking activity.

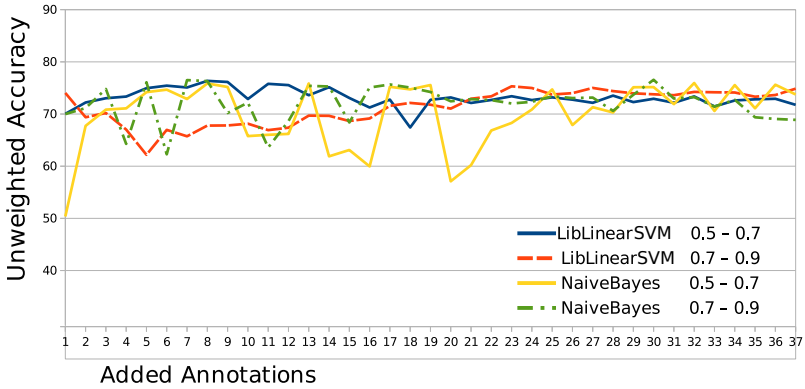


Figure 6.10: Training progression using different confidences and models

6.3.3.5 Active Learning

Our Active Learning implementation uses query on uncertainty for sample selection, see Section 6.2.2 for further background. The credibility range that triggers user requests can be specified, thus DrinkWatch supports relevance sampling as well as uncertainty sampling. The option is part of an online classifier component shown in Figure 6.8. It manages the assembly of sample lists from user annotations and data streams as well as the training process of our online model. The model's predictions are also handled by the online classifier. Both, requests and predictions, are handled as events instead of streams with fixed sample rate.

6.4 Evaluation and Results

Following system implementation and data collection, three steps of evaluation are presented in this section: the static evaluation of the fully annotated data set in Section 6.4.1, the evaluation of different learning strategies in Section 6.4.2, and the interactive run performed with end users in Section 6.4.3.

6.4.1 Evaluation of Static Models

To give an overview of the collected data and provide an impression of what accuracy fully trained models are able to achieve, Table 6.1 shows results of Naive Bayes and linear SVM (implementation: LibLinear) models trained on the full data set, evaluated on the fixed test set that is also used for the simulation of cooperative ML.

	Results of full Training	
	Naive Bayes	**linear SVM**
Drinking	81.4%	84.9%
Not drinking	71.6%	79.8%
Unweighted Average	**76.5%**	**82.3%**

Table 6.1: Results of training on all annotations contained in the training set, evaluated on the test set.

Our results are in line with other results on drink activity recognition found in literature, where drink activity is can be classified on accelerometer data with 70 % to 80 % accuracy [226]. The linear SVM model shows a six percent points lead over Naive Bayes, which again is as expected. While there is a difference on the ”Drinking” class, the difference is larger on the ”Not drinking” class. As ”Not drinking” is by far larger and more complex, Naive Bayes struggles in finding a model, that depicts the classes’ behavior in recorded data.

6.4.2 Learning Strategy Simulation

Since it is intended to utilize the learning process within an end user application, that is designed to continuously adapt to the specific activity patterns of the user, it makes sense to not only evaluate the complete model, but also the relative improvements of the classifier when increasing the amount of training data. To evaluate this continuous refinement of the classification system, the iterative training process is simulated by using the NOVA [17] toolkit.

First of all, the base model is trained on a small stack of eight annotations from one session. From there on this is used as baseline classifier to predict the rest of the training data. Subsequently, the first label where the confidence is equal or greater than the lower end of a predefined confidence interval is selected. In case the confidence value lies within the interval the oracle is queried to correct the answer. The oracle is simulated by the full hand-labeled annotation. In case the confidence value of the prediction is higher than the upper limit of the interval it is assumed that the classification of the sample is correct, and forwarded to the logging application. Afterwards the newly annotated sample is added to the training data and the classifier is retrained before repeating the same steps again. This process continues iteratively until all available data has been annotated. While in theory the classifier could learn from data

with high classification confidence and improve without explicit user input, the system sticks to user (oracle) labeled data only because those are guaranteed to be true positive samples as long as the user gives correct feedback.

The study has been conducted by applying an uncertainty sampling strategy which utilizes a low confidence interval ranging from 0.5 to 0.7 as well as a relevance sampling strategy using a high confidence interval from 0.7 to 0.9, see Figure 6.10. While one would expect the unweighted average accuracy to increase steadily with the number of available training data our simulation results paints a different picture as shown in Figure 6.10. Naive Bayes is clearly more unstable than LibLinear's linear SVM. Obviously, it is less robust against variations across sessions and users as well as untypical drink activities, for example, those with long pauses while holding the vessel.

All models stabilize over the course of the simulation. By the time 30 additional labels are added to the base stock, the variations in accuracy narrow down to five percent points for Naive Bayes and three percent points for LibLinear, when adding new labels to the training process. While low confidences seem to be preferred by the LibLinear SVM model, queries based on high confidences seem to be the better choice for Naive Bayes. The progress of both models is best judged using predictions, as shown in Figure 6.9. One can see where the classifier triggers and with what confidence, as indicated by hatching and color. The first line contains the hand labeled annotation and is followed by predictions of models with increased number of training data. Areas marked in green are drinking activity. Naive Bayes changes in accuracy, seen in Figure 6.10 manifest themselves as low confidence, red bars on the right.

6.4.3 Interactive Machine Learning Sessions involving Bodystorming

Bodystorming [141, 164] typically is associated with early stages in the creation of embodied interaction design "in the wild". Since DrinkWatch is targeting natural behavior of drinking as means of interaction with a smart watch, Bodystorming is in this case used as evaluation method. In a natural setting users are confronted with DrinkWatch to describe their experience of body motion and system feedback. Thus, aspects of the users' motion, described aloud can be gathered together with system behavior. The feedback can be seen as final step of the first iteration in development of our interactive machine learning system based on hand motion. Adoption of Bodystorming as evaluation is a step towards employing interactive machine learning for movement interaction design as sketched by Gilles et al. [80].

Two users were invited to use DrinkWatch for one hour to track their drink activities. For this experiment, the high confidence range is picked (0.7 to 0.9) since it promises an earlier stabilization for Naive Bayes. To create a reliable base model, at least 40 annotations were used. The users were able to judge the quality of the model by the appropriateness and frequency of

queries. While both users had the impression that drink activities were accurately recognized in general (e.g. “Five out of six” stated by one of the two users), there were many wrong positives due to the unbalanced nature of both classes. The unfiltered requests were described as annoying by the users and made the system unusable. The behavior of the system appeared transparent to users. They noted that moving a vessel containing fluid, slow and steady was a key trigger for recognizing drink activities. It was also easy for them to mimic activities triggering the model, describing properties of the movement that lead to requests.

6.4.4 Discussion

The need for mobile interactive and cooperative ML approaches is motivated by shortcomings of classical ML approaches, considering (i) difficulties in getting authentic data of every day living, and (ii) a deficit of transparency and user control. Interactively integrating users into the ML process would have the potential to address both issues, allowing users to label their own activities, to gain some understanding of and control over machine functionalities, and to ultimately peek behind the curtain of automation and to leave users with a feeling of competence and self-efficacy.

Since mobile cooperative learning is a novel research area with many conceptually and technically open issues, this chapter exposed the process of developing the DrinkWatch application and its integration with smart data sources, such as the smartscale. The intention and aim was to become able to infer limitations and potentials of future mobile cooperative ML application. After developing the core functionalities of the DrinkWatch application, a time period of six months followed, iterating the application based on multiple tests, including a longer period of time testing the application with myself and short episodes collecting insights from letting colleagues and friends try the application. Building and testing DrinkWatch, showed that interactive cooperative machine learning is already feasible on today’s state of the art smart watches. Feedback provided by the model (i.e. the machine intelligence) as a direct consequence to a drink activity is intuitively graspable by users, as think-aloud sessions indicated, even when feedback is provided through simple audio notifications. Based on the model performance in recognizing drink activities, it can be assumed (as it is typical with many ML based models) that it can be adopted easily to recognize other hand-based activities.

LibLinear’s linear SVM does not only show higher accuracy compared to Naive Bayes, but also a smoother learning curve. Since both models have opposing tendencies when it comes to confidence intervals, fusing both models in a Query by Committee [193] implementation, seems promising. The committee might also be accompanied by static models, such as the warmstart model or save points that can be created by the user as well.

As also described in the interactive ML paradigm [5], queries should be forwarded to the user with care, since wrong positives cause frustration and users tend to describe the experience

associated with wrong positives as “annoying”. When it comes to adaptability to new health related hand activities, this chapter presented several observations that can be used as reference points. The minimal strength of the Naive Bayes warmstart model for the problem of drink activity recognition can be set at circa 40 overall annotations, this equals about eight hours of recording in this case while the linear SVM stabilizes at about 30 overall annotations or six hours of recording.

The smart scale was introduced as an option to integrate data from other data sources, in the hope to improve the (initial) quality of the model. The use of a smart scale reduced the annotation effort drastically and helped to understand that it is a suitable physical object to facilitate logging of fluid intake as well as to support annotation and online learning on smart watches in a stationary setting. Interactive machine learning on mobile devices also was researched with respect to explainability in [70], that goes a step further in visualization of the process of a machine learning classification on images. Here, saliency maps of a deep neural network were calculated on images of aesthetic and not aesthetic forest. Moreover, a study was presented that underlined that saliency maps increase transparency of the machine learning process and help the user to sharpen his view regarding images of forest scenery.

6.5 Summary

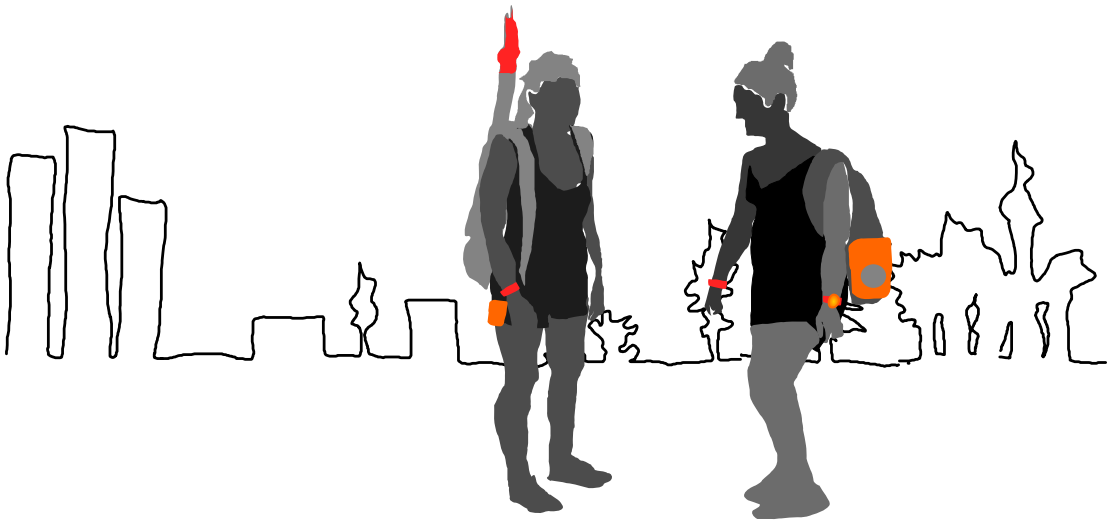
The design and evaluation of DrinkWatch, a smart watch application for drink activity recognition, showed possibilities and challenges when it comes to the method and the technology of mobile interactive machine learning systems. A key motivation for the development of a mobile system employing interactive machine learning, lies in the circumstance, that it is not feasible to annotate data of recorded wrist motion in retrospective. Instead annotation has to take place at the time an action is executed. With DrinkWatch, the effort in annotation can be shared between user and smart objects. Moreover, active learning is implemented to select only data that are the most important for the machine learning process for labeling.

Since machine learning models can be trained incrementally without the need for storing data, the user can be provided with a more reactive system that improves his privacy at the same time. For this purpose, SVM and Naive Bayes were evaluated with simulated strategies of different confidence ranges. Here data recorded using MobileSSI in the wild were used as a basis.

As a further evaluation step, bodystorming was conducted with users and the DrinkWatch prototype. Bodystorming lies an emphasis on embodied experience and is usually employed early in the design process. First experience of users with the prototype lead to the insight that users were able to identify properties of their hand motion the classifier reacted to. They noted that moving a vessel containing fluid, slow and steady was a key trigger for recognizing drink activities.

Chapter 7.

Mobile Recognition of Wellbeing within Local Climate Zones



This Chapter is based on the publication *Mobile Sensing for Wellbeing Estimation of Urban Green using Physiological Signals* [69]. Own contributions lie in developing the software setup for data collection, creating mobile user interfaces and evaluation via machine learning. Study design was realized with the help of Christoph Beck and Joachim Rathmann working at the Institute of Geography at the University of Augsburg, while Andreas Seiderer contributed custom hardware. In addition to the evaluation presented in the paper, relying only on blood volume pressure, evaluation of further sensors (skin conductance, audio) and fusion of physiological sensors (blood volume pressure and skin conductance) was integrated in this chapter.

7.1 Conception of mobile label acquisition and context recognition

In Chapter 5, environment with their different noise levels are identified as challenges in MSSP. Beyond auditive characteristics there are further influences thus as heat and humidity that have

influence on us, depending on where we are. When viewed from the perspective of wellbeing, environments can bring a multitude of positive factors, such as an incentive to exercise one's body in walkability [74] or positive effects on recreation.

Thus, this chapter tackles environment, foremost in regard to personalized models of wellbeing, depending on local climate zones. For this purpose study participants were exposed to three different local climate zones, while a variety of sensors was recorded synchronized with the participant's self-assessed rating of wellbeing (valence). Later on machine learning was used to classify on the one hand the local climate zone and on the other hand the self-assessed valence based on recorded sensor data.

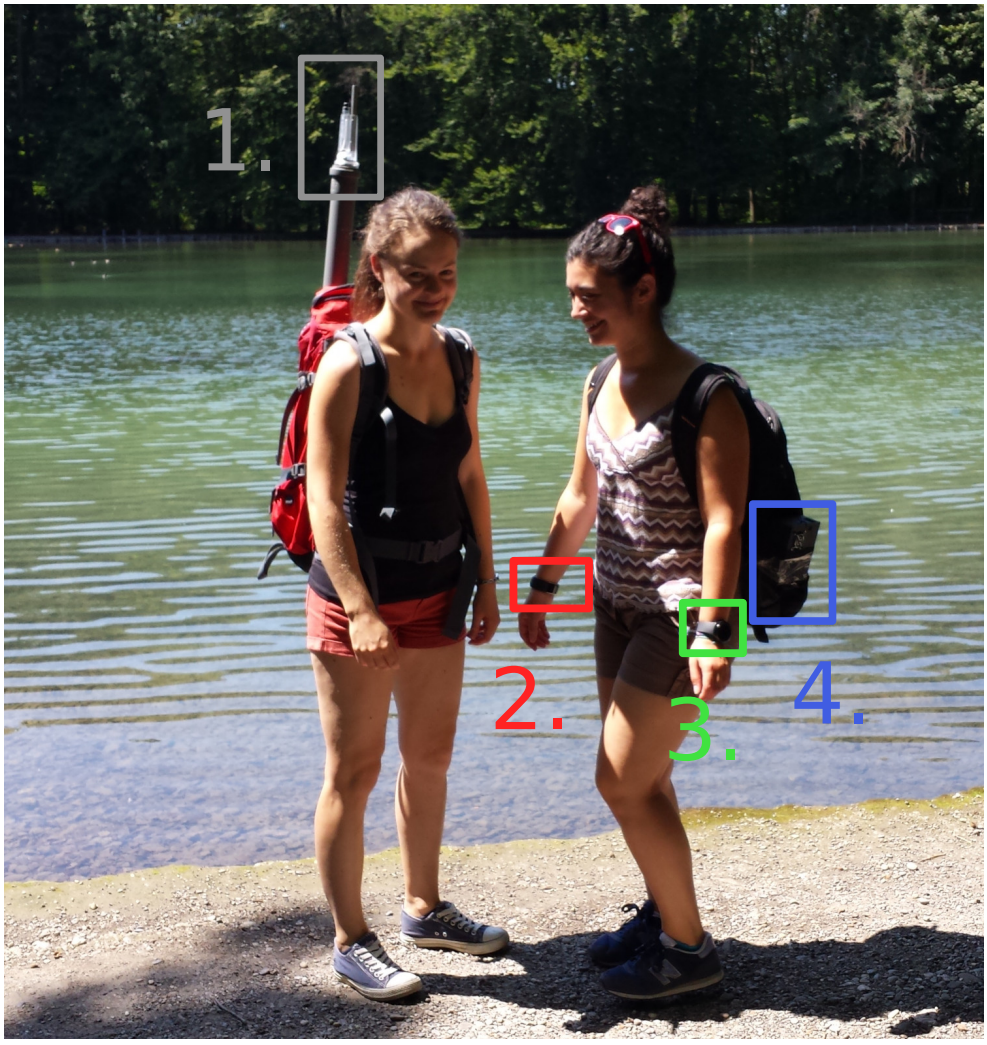


Figure 7.1: Two participants taking part in the field study. The following sensor devices are visible in the photo: 1. Aspiration psychrometer, 2. Microsoft Band 2, 3. Samsung Gear S2, 4. Custom built Environmental sensor box

Nature therapies, such as garden therapy [1] or Shinrin-yoku forest bathing [145], have provided evidence of the positive effects of green on humans' mind and physique [160]. Previous research argues for a healthy effect of natural environments even when only viewed through

a window [212]. Going beyond the typical quantified-self notion ”know thyself”, this chapter aims to employ sensory data to investigate the potential of health-transpiring environments.

There is a large body of work that employs HCI technology to enhance interaction with nature (e.g., [25]), including explorations of human-plant interactions. However, understanding what aspects of an environment influence wellbeing and integrating this knowledge dynamically into (technology) design is a challenging task. Relevant data to analyse a person's wellbeing may be acquired, for example, with mobile and stationary sensors. Since sensor other data other than video or audio do not speak for themselves when viewed out of context, it is essential to request specific annotations from the user in the very moment of data recording [50]. To this end, comfortable mobile interfaces have to be provided that facilitate the input of annotations without affecting the user's experience of the outer-body environment. A number of applications draw on the interdependencies between the aesthetics of a landscape and an individual's wellbeing. Examples include navigation systems that aim to reduce environment-induced stress on the user. Contemporary approaches consider routes with beautiful scenery instead of the fastest route [173] when generating recommendations. Such applications go beyond routing and can even recommend on which side of a bus [179] passengers should sit on a bus tour to experience the most aesthetic views during their trip. Usually, these applications rely on machine learning models that are trained using image data from the categories of interest, i.e. aesthetic and non-aesthetic scenery, such as a highway. However, they do not employ objective physiological or behavioral measures to assess the user's wellbeing.

A variety of wellbeing models has been proposed in the literature [56] to capture relevant factors, such as sufficient sleep and healthy nutrition, that influence an individual's wellbeing. Besides subjective measures, usually drawing on valence-based self-reports, objective measures, such as physiological data, are employed to assess an individual's wellbeing [98, 100]. Also, the influence of an individual's environment on physiology has been researched by measuring heart rate and heart rate variability [200], skin temperature and pulse rate [133]. Climate is a key environmental influence on the human body, therefore it plays a central role in this study design. Fine grain assessment on environmental and personal factors combined so far have not been used together with machine learning to gain models that can be personalized. Nonetheless, adaption of machine learning models on mobile device is a feasible task [188]. While deep learning is becoming increasingly popular in Affective Computing [227] this chapter relies upon handcrafted features due to the sparse nature of the data set [223]. This leads to the main question to be answered with following work: How can the influence between the users' urban environment and their physiological and psychological wellbeing be modeled? To shed light on this question, a field study with 20 participants is conducted. Measurements were recorded with a wide range of sensors associated with the participants' inner-body states and their urban environment. In addition, self-reports of psychological parameters are recorded, including valence-based ratings of the pleasantness of the moment as well as subjective assess-

ments of thermal sensation and perceived air quality. Due to different screen sizes of smart phone and smart watches, the users are provided with dedicated graphical user interfaces for each of these devices, to conduct the labeling. Since people are usually able to distinguish easily between a pleasant and an unpleasant feeling [14], it is possible to annotate data related to well-being on the go. In the following, the setup and results of a study are presented to investigate dependencies between the urban environment and the user's wellbeing as a first step towards an application fostering the user's wellbeing while employing an interactive machine learning subsystem [68, 167].

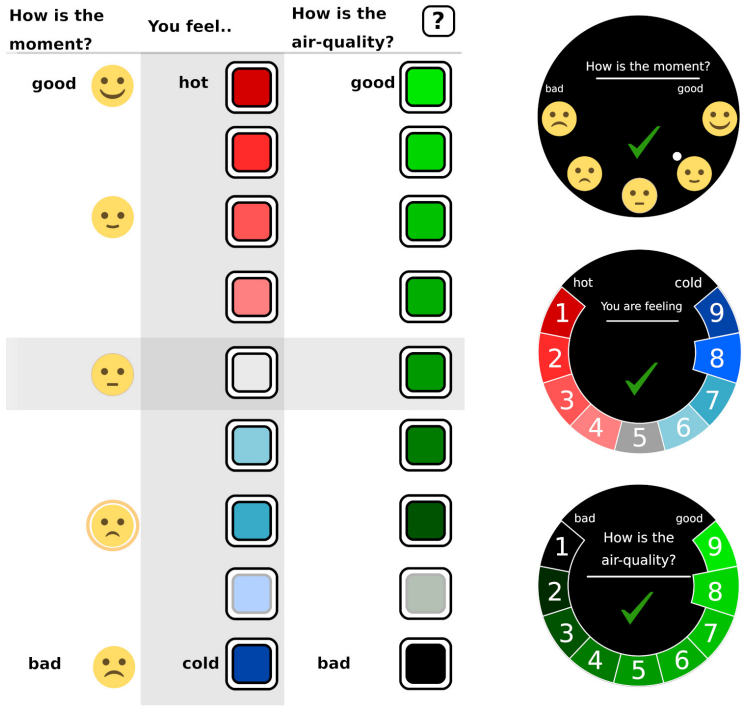


Figure 7.2: GUI used for self-assessment on smart phone (left) and smart watch (right). Each version gathers wellbeing in terms of valence (5 point scale), subjective impression of temperature and air quality (9 point scales).

7.2 Setup and Data

In each recording session, data was collected from two participants, with one participant using a smart watch for labeling, and the other using a smart phone. The user interfaces for both devices are shown in Figure 7.2. In addition to the annotation devices, both participants wore fitness bands to collect physiological data. The per-person setups also varied in the addition of an aspiration psychrometer, for the detection of ambient air temperature, relative humidity and further derived variables relevant for wellbeing. Aspiration psychrometers are bulky professional devices, the established standard for measuring relative humidity. They serve as high-quality reference for low cost alternatives integrated into the custom-built sensor box as displayed in Figure 7.1.

Figure 7.3 shows the setup for acquistion of sensor data and users' self ratings. Different sensors tend to provide data at different speeds, see Table 7.1. Especially, on mobile devices, chunks of data at one moment would span different time periods for different sensors. Therefore synchronization of the individual data streams is required. To record data, mobile tools for the acquisition and analysis of sensory data are employed, SSJ [44] and MobileSSI [66], which include specific mechanisms for synchronizing multi-sensor data.

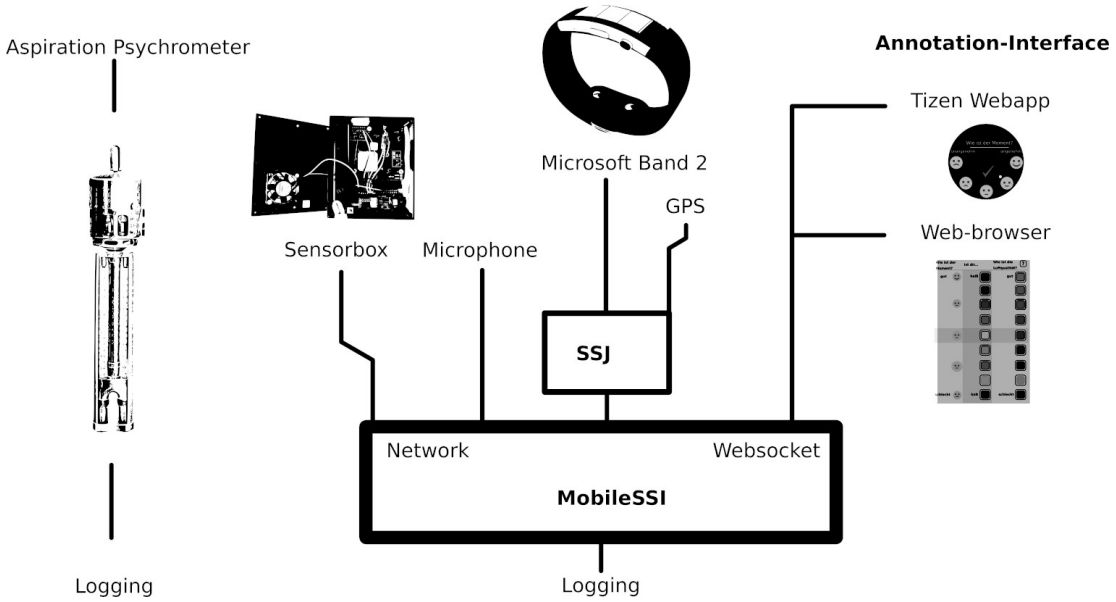


Figure 7.3: Setup including sensor configuration, recording software and annotation interfaces.

7.2.1 Sensors and Devices

While the evaluation of this chapter focuses on physiological signals, the recording setup was done using a wide range of sensors that promise useful input for context-aware applications. Environment-related data were collected using microphones connected to the smart phones, GPS provided by the smart phones, and a custom-built sensor box. The sensor box contained a sensor for temperature, humidity (SHT75) and air pressure measurement (BMP280) as well as dust (SDS011) and gas (MICS) sensors. For a detailed overview, see Table 7.1. Person-related data were collected via a Microsoft Band 2 providing galvanic skin conductance (GSR), heart-rate (HR) and the heart's inter-beat-interval (IBI) as physiological signals. Aspiration psychrometers were used to collect temperature and humidity as an indicator of heat stress. These devices only store data locally, thus synchronization was required which was conducted retrospectively via GPS and timestamp information.

Device	Sensor	Data	SR (Hz)
MS Band 2	BVP sensor	IBI	30
		HR	1
	GSR	GSR	5
	Accelerometer	ACC	62.5
Smart phone	GPS	coordinates	5
	Microphone	audio (raw)	16000
	UI	ratings	event based
Smart watch	UI	ratings	event based
Sensor Box	SDS011	PM2.5	0.07
		PM10	0.07
	SHT75	humidity	0.07
		temperature	0.07
	MICS	CO	0.07
		NO_2	0.07
		NH_3	0.07
		C_3H_8	0.07
		C_4H_{10}	0.07
		CH_4	0.07
		H_2	0.07
		C_2H_5OH	0.07
	BMP280	pressure	0.07
		temperature	0.07
Aspiration Psychrometer		humidity	0.5
		GPS	0.5
		temperature	0.5

Table 7.1: Devices involved in the recording setup.

7.2.2 Experience Samples - Label Acquisition

In order to gain information on the participants' wellbeing, they are asked to provide explicit experience samples about their momentary state, related to valence (5-point scale, good to bad), perceived temperature (9-point scale, hot to cold) and air quality (9-point scale, good to bad).

On the back-end side, annotations were serialized synchronously with the collected data described in Section 7.2.1. Participants were asked to annotate whenever they felt a change of the respective states. Thus, a label was valid until a new label replaced it. Overall 769 labels where annotated on the go by the study's participants with an average of 38 labels per session. Since those samples represent impressions for a short time period only, questionnaires including ratings related to mean and variance of wellbeing, temperature and air quality concluded each session.

7.2.3 Route and Sessions

The route used for recording was selected due to the variety of local climate zones including the built up "Open Mid Rise" as well as the mainly natural "Scattered Trees" and "Dense Trees" categories [19, 202] (see Figure 7.6). These local climate zones covered both exposure to heat

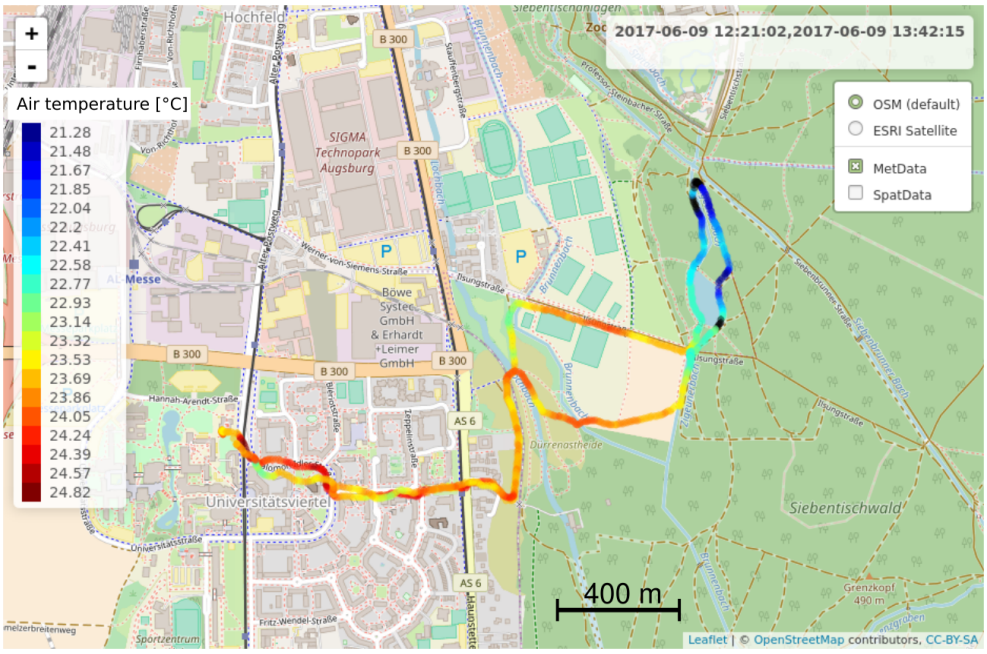


Figure 7.4: Plot of temperature acquired along the route in one exemplary measurement.

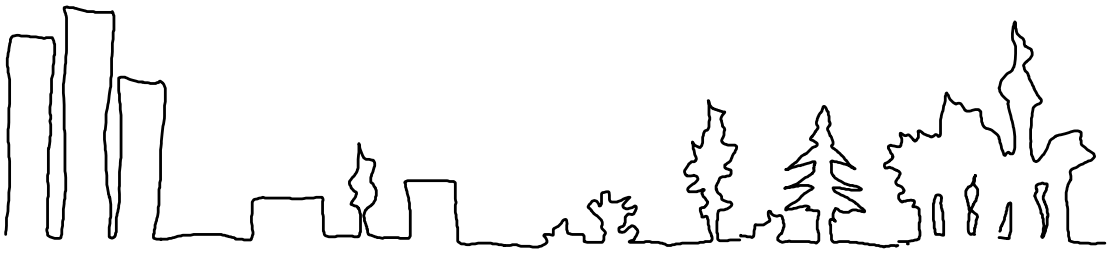


Figure 7.5: compact high-rise, open midrise, low plants, scattered trees, dense trees

in city and open meadow as well as sheltering forest. In addition to the GPS-track marking the route, Figure 7.4 also shows the temperature along the track for one exemplary walk. Each session took about 80 minutes for the 5 km walk. Days with similar weather conditions were chosen for the study in order to reduce a weather-based bias on data. Measurement campaigns were performed under clear and calm weather conditions around noon on midsummer days to ensure maximum thermal differences between local climate zones and thus increasing the chance of capturing potential heat stress. The path used for data-acquisition is conceptualized as a loop, to prevent the participant's time walking from influencing physiological data. A total of 20 sessions with 7 participants were included in the data set used for the evaluation in Section 7.3.

7.3 Evaluation and Machine Learning Models

To cope with the complexity of the data, machine learning techniques are utilized, while searching for meaningful patterns across physiological data (BVP), characteristics of the environment

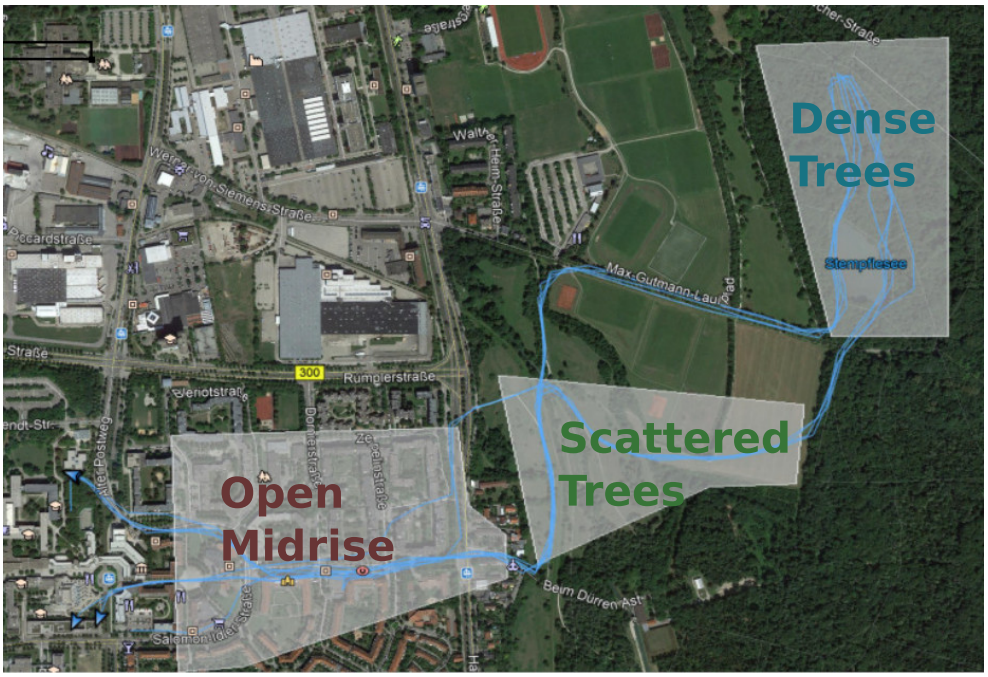


Figure 7.6: Occurring local climate zone types: Open Mid Rise (City), Scattered Trees (Meadow), Dense Trees (Forest).

and the users' self-reported wellbeing. Since the environment is a key independent variable in this study's design, the influence of environment on physiology is tested by discriminating different climate zones using physiological data. An important objective of the taken approach is to investigate to what extent the user's experience in terms of wellbeing, temperature and air quality may be predicted from the recorded physiological data. The user's subjective ratings for pleasantness of the moment, perceived temperature and air quality were used as a golden standard in the training and evaluation process. Once a model is trained, it can directly be employed to generate context information for an application in real-time [66]. In addition, the ability to process data directly on the device is an important prerequisite to respect the user's privacy, that again is a desirable feature when processing health related, physiological data. In the following the steps involved in the data processing pipeline are described.

7.3.1 Feature set for Machine Learning

In order to calculate features for the classification approach, data provided by the Microsoft Band 2 are used for physiological data, which provides inter-beat-interval (IBI) and heart rate (HR) and Galvanic Skin Response (GSR) that are analyzed in the following. Usually, physiologically measurable effects of a change in the state of mind (such as an spontaneous increase in valence) are slightly delayed and rather long-lasting. While the Microsoft band provides raw data for GSR, the Blood Volume Pressure (BVP) is already processed into HR and IBI (see Figure 7.7). Therefore, a range of statistical features is applied over multiple data samples to cover the

characteristics of the input signal over an extended time period.

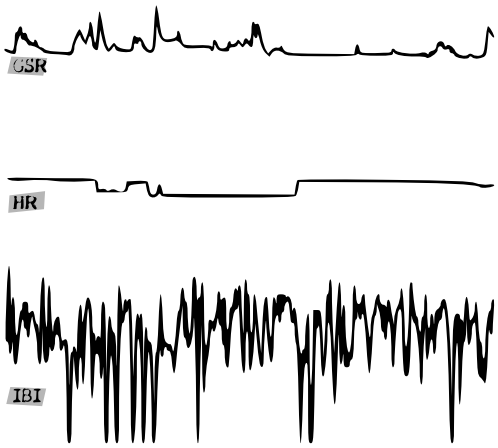


Figure 7.7: Physiological Data: GSR, HR and IBI

To assess the contribution of each statistical feature on HR and IBI to the overall classification result, a Sequential Forward Selection (SFS) was applied. The results of this SFS are shown in Table 7.2. SFS identifies the most useful single feature to which the next best feature is added subsequently. Recognition rate at rank 2 therefore is achieved using feature 1 and 2. The full feature set is used for the training of models presented in the following section.

In Table 7.2, the score peaks the first time at rank 9 with a score of 0.5. A closer look at the results of the SFS reveals that mostly IBI-Features are used to reach peak performance, which indicates that the heart's inter-beat-interval conveys more valuable information than the heart rate in our case. HR (1.0 Hz) and IBI (5.0 Hz) features are calculated on the same time slice, consisting of a 10 second frame containing new data, and a 240 to 380 seconds overlap containing old data.

rank	feature	score	rank	feature	score
1	IBI_MAXPOS	0.40	12	HR_ZEROS	0.50
2	HR_MAXPOS	0.43	13	IBI_PEAKS	0.50
3	IBI_MIN	0.43	14	IBI_MINPOS	0.49
4	IBI_STD	0.43	15	HR_STD	0.41
5	HR_PEAKS	0.43	16	HR_RANGE	0.35
6	IBI_ZEROS	0.43	17	IBI_LEN	0.34
7	IBI_MAX	0.42	18	HR_LEN	0.34
8	IBI_ENERGY	0.45	19	HR_MEAN	0.33
9	HR_MINPOS	0.50	20	HR_ENERGY	0.33
10	IBI_MEAN	0.48	21	HR_MAX	0.34
11	IBI_RANGE	0.50	22	HR_MIN	0.31

Table 7.2: Sequential Forward Selection (SFS) of 22 BVP related Features.

In addition to features concerning HR and IBI, features on GSR are used. Since the the most valuable information in GSR is not on the overall value, but in the signals characteristica such as peaks and slopes, those events are the foundation of the employed GSR-feature-set. Additionally area, amplitude and duration of each event are calculated that again are fed into five functionals, see Figure 7.8.

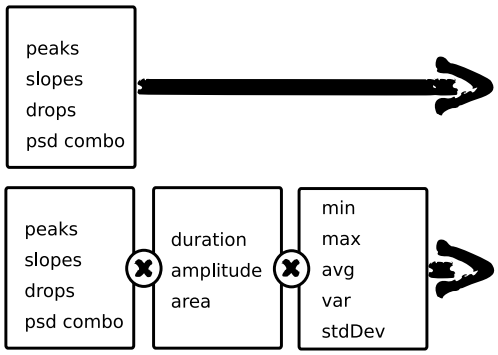


Figure 7.8: 64 GSR Features, based on peaks, slopes and drops

Next to person related, physiological data, audio data are used for classification. The audio features are based on MobileNet [89]. Audio data are converted into images of mel-spectrograms, that are subsequently put into MobileNet to produce 1792 features that represent MobileNet's last layer before reduction to the target class-count takes place. The ANN is so to speak cut, to gain a feature extractor. Using spectrograms with network architectures for image based recognition is known to perform well on classification on audio data [6].

7.3.2 Machine Learning Models

The models described in this section were evaluated using cross-validation, see Table 7.3 to Table 7.15 for details. As a classificator we chose Support Vector Machines (SVM). Since SVM's are sensitive to unbalanced sample distributions across classes, random undersampling was used. An important factor for real-time applications is the responsiveness. Therefore, the frame sizes are kept as low as possible, which enables classification results at a higher frequency and thus increases the responsiveness of the system.

7.3.2.1 Environmental Context on Physiological Data

In the first experiment it is investigated, to what extent it is possible to infer the outer-body environmental context from the recorded physiological data. This outer-body environmental context is modeled by means of the landscape the user is currently situated in. Consequently, it is the goal of the model to distinguish between the three classes *dense trees*, *scattered trees*, and *open midrise*, representing the according local climate zones. This helps us counter-check if there is an influence of the environment on the participants' physiology. The labels were gained by manually defining the different landscapes on the route and automatically establishing a mapping with the GPS coordinates of the user. This mapping is also shown in Figure 7.6.

Heart Rate and Inter Beat Interval

BVP sensors are common in consumer fitness wearables and therefore a desirable modality for measuring the influence of environments onto wellbeing "in the wild".

	Environment HR & IBI (SVM)			
	dense trees	**scattered trees**	**open midrise**	**Acc. %**
dense trees:	132	16	12	82.50 %
scattered trees:	10	61	89	38.12 %
open midrise:	12	87	61	38.12 %
Average				**52.92 %**

Table 7.3: User related model trained over 945 samples, 10 seconds frame, 240 seconds overlap, 2-fold cross-validation

Results on the classification of the local climate zones chosen in the original experiment design, shows that *dense trees* can be recognized well while *scattered trees* and *open midrise* are often confused.

	Environment reduced HR & IBI (SVM)		
	dense trees	**open midrise & scattered trees**	**Acc. %**
dense trees:	167	50	76.96 %
open midrise & scattered trees:	77	140	64.52 %
Average			**70.74 %**

Table 7.4: User related model trained over 434 samples, 10 seconds frame, 240 seconds overlap, 10-fold cross validation.

This leads to reducing the target classes from three to two, raising the classification results drastically.

Galvanic Skin Response

GSR is tested as second physiological modality. It is trained on the same frame-sizes as BVP-related data and results in a higher accuracy that could be achieved on BVP even with class reduction.

Here classification of *open midrise* is best, *scattered trees* second best and classification of *dense trees* is lower at 70.62 % than when relying on BVP-related data, where *dense trees* is classified with 82.50% accuracy.

	Environment GSR (SVM)			
	dense trees	**scattered trees**	**open midrise**	**Acc. %**
dense trees:	113	19	28	70.62 %
scattered trees:	18	118	24	73.75 %
open midrise:	16	16	128	80.00 %
Average				**74.79 %**

Table 7.5: User related model trained over 480 samples, 10 seconds frame, 240 seconds overlap, 2-fold cross validation

Feature Fusion

Consequently fusion of both modalities is promising. Since both BVP and GSR related data are typically giving best results in the same time-window, feature fusion is employed, by merging the individual feature vectors.

	Environment Fusion (SVM)			
	dense trees	**scattered trees**	**open midrise**	**Acc. %**
dense trees:	129	10	21	80.62 %
scattered trees:	9	148	3	92.50 %
open midrise:	9	4	147	91.88 %
Average				**88.33 %**

Table 7.6: User related model evaluated on 480 samples, 10 seconds frame, 240 seconds overlap, 2-fold cross validation

This results in another mayor boost, reaching 88 % accuracy on three classes. It can be noted that *dense trees* are still classified worse than they would be with a BVP-based model, which is why reducing class-count from three to two is reconsidered.

	Environment reduced Fusion (SVM)		
	dense trees	**open midrise & scattered trees**	**Acc. %**
dense trees:	203	14	93.55 %
open midrise & scattered trees:	16	201	92.63 %
Average			**93.09 %**

Table 7.7: User related model evaluated on 434 samples, 10 seconds frame, 240 seconds overlap, 10-fold cross validation.

combining *open midrise* and *scattered trees* leads to an additional improvement to over 90 % accuracy. Environments, especially viewed as forest and and not forest, have an impact on the human body that can be automatically recognized via consumer grade hardware and mobile machine learning.

7.3.3 Environment Related Wellbeing on Audio Data

The individual's wellbeing relying on self-assessment is an additional key question that is tried to be answered with the experiment-setup. While the environment's state was classified above using the measured body state, classifying how *good* or *bad* a person feels in five discrete steps is firstly examined relying on audio data.

The LibLinear SVM classifier is trained on a feature vector of size 1792, that is extracted from MobileNet V2 [180]. MobileNet V2 is selected based on evaluation done by Seiderer et al. [188] for Convolutional Neural Networks' performance on mobile devices. Since environment classification is different from classification of paralinguistic phenomena such as laughter in Chapter 5, a feature-set different from EmoVoice [219] is selected. While specialized architectures exist [9] their performance is inferior to the application of MobileNet on spectrograms in our tests. The use of SVMs increases the flexibility and speed of the learning progress compared to only using a Deep Neural Network.

Regarding data quality, audio data proved to be not as robust as physiological data. Due missing and unusable data only four sessions remained for model creation and evaluation. While there is a considerable amount of talking next to environmental sounds such as birds on the recording, cutting out sections with human voice did not lead to a change in model quality.

Valence on Audio (SVM)

		1	2	3	4	5	Acc. %
1	**(good)**	172	32	26	20	18	64.18%
2		11	174	53	24	6	64.93%
3		15	61	158	32	2	58.96%
4		23	21	35	141	48	52.61%
5	**(bad)**	18	19	14	34	183	68.28%
Average							**61.79,%**

Table 7.8: User related model evaluated on 1340 samples of audio data with 4.13 seconds frame without overlap using 10-fold cross validation.

This leads to over 60% accuracy, which is clearly over chance. The classes "1", "2" and "5" score best, as can be seen in Table 7.8, with class "4" scoring worst.

Reducing classes from five to three improves the results, since the newly created extreme classes are recognized well, whereas the neutral class stays unchanged, compare Table 7.9.

7.3.4 Environment Related Wellbeing on Physiological Data

It is more common to classify a persons wellbeing by measuring the body-state directly instead of relying on modalities that describe his context.

Valence on Audio (SVM)					
		1 & 2	**3**	**4 & 5**	**Acc. %**
1 & 2	**(good)**	86	7	2	90.53 %
3		20	56	19	58.95 %
4 & 5	**(bad)**	5	22	68	70.58 %
Average					**73.68 %**

Table 7.9: User related model evaluated on 285 samples audio data, with 4.13 seconds frame, without overlap using 10-fold cross validation.

Heart Rate and Inter Beat Interval

Therefore physiological signals are considered in the following.

Valence on HR & IBI (SVM)							
		1	**2**	**3**	**4**	**5**	**Acc. %**
1	**(good)**	575	56	96	92	40	66.94 %
2		40	541	199	89	55	62.98 %
3		147	172	148	179	213	17.23 %
4		114	149	120	307	169	35.74 %
5	**(bad)**	137	104	109	177	332	38.65 %
Average							**44.31 %**

Table 7.10: User related model evaluated on 4295 samples, 10 seconds frame, 240 seconds overlap, 10 fold cross validation.

A model based on BVP-related features scores lower than the Audio related model, while classifying the neutral ”3” class lower than chance at 17 % whereas classes ”1” and ”2” score best at over 60 % accuracy.

Valence on HR & IBI (SVM)					
		1 & 2	**3**	**4 & 5**	**Acc. %**
1 & 2	**(good)**	214	26	57	72.05 %
3		41	205	51	69.02 %
4 & 5	**(bad)**	102	84	111	37.37 %
Average					**59.48 %**

Table 7.11: User related mod elevaluated on 891 samples, 10 seconds frame, 240 seconds overlap, 10-fold cross validation.

Reducing classes to three, improves the results for the *good* (”1&2”) and *neutral* class (”3”) whereas the *bad* class scores worse, leading to an overall improved result.

Galvanic Skin Response

Considering GSR only, classification results are better than those achieved on audio data with 65.53% on the five-class problem and 75.08 % on the three-class problem.

Valence on GSR (SVM)

		1	2	3	4	5	Acc. %
1	**(good)**	141	5	16	13	24	70.85 %
2		15	150	15	9	10	75.38 %
3		18	17	142	2	20	71.36 %
4		23	23	8	107	38	53.77 %
5	**(bad)**	27	22	13	25	112	56.28 %
Average							**65.53 %**

Table 7.12: User related model evaluated on 995 samples, 10 seconds frame, 240 seconds overlap, 10-fold cross validation.

Valence on GSR (SVM)

		1 & 2	3	4 & 5	Acc. %
1 & 2	**(good)**	245	16	40	81.40 %
3		13	226	62	75.08 %
4 & 5	**(bad)**	43	51	207	68.77 %
Average					**75.08 %**

Table 7.13: User related model evaluated on 903 samples, 10 seconds frame, 240 seconds overlap, 10-fold cross validation.

Classes *"4" & "5"* score worst, and the gap is reduced only slightly compared to neutral and *good* classes with class-reduction.

Feature Fusion

Since Feature fusion helped drastically on the problem of environment classification, it is employed here also, even though both modalities are the weakest on *bad* self-assessed wellbeing.

Where feature fusion boosts results, to 79.68% from 75.08% the leap is not as big as it was in classifying environments on physiological data. Reducing classes leads to a model that is more evenly strong on its individual classes compared to the five-class model. That apart, improvements are sightly.

Valence on Feature Fusion (SVM)							
		1	**2**	**3**	**4**	**5**	**Acc. %**
1	**(good)**	176	0	7	3	0	94.62 %
2		0	170	2	5	9	91.40 %
3		6	9	163	3	5	87.63 %
4		7	21	13	100	45	53.76 %
5	**(bad)**	9	4	13	28	132	70.97 %
Average							**79.68 %**

Table 7.14: User related model evaluated on 930 samples, 10 seconds frame, 240 seconds overlap, 10-fold cross-validation.

Valence on Feature Fusion (SVM)					
		1 & 2	**3**	**4 & 5**	**Acc. %**
1 & 2	**(good)**	257	12	28	86.53 %
3		7	254	36	85.52 %
4 & 5	**(bad)**	41	43	213	71.72 %
Average					**81.26 %**

Table 7.15: User related evaluated on 891 samples, 10 seconds frame, 240 seconds overlap, 10-fold cross validation.

7.4 Discussion

This chapter investigated different modalities and their fusion in respect to environment related wellbeing. Further work to a personalized approach to the research of the recreational aspects of forests has been presented [70], where aesthetic scenery is classified on the go.

Audio data of soundscapes as well as physiological signals were evaluated in respect to classification of users' self-assessed wellbeing and environmental classes. Thus, a mutual influence between body state and environment can be outlined. Starting at widely available BVP sensors' data such as heart rate, adding Skin Conductance (GSR) in the progress. The relation between body and environment is approached by classifying physiological data under two annotation schemes. Firstly considering local climate zones (LCZ) derived from GPS and secondly referring to participants' self-assessed wellbeing.

Considering the concrete classification results on LCZ, *dense trees* can be classified well (82.50 %) on BVP related data only, while overall classification results score low at 52.92 %.

GSR-data yield better results at 74.79% unweighted accuracy, while achieving best results on the *open midrise* class.

Feature fusion of GSR and HR/IBI improves classification results by over 13 percent points to *88.33%*, class reduction, restricting the problem complexity to classifying *dense trees* or $\neg$*dense trees* (*open midrise & scattered trees*) improves the results again to 93.09%.

This suggests that physiological data from consumer grade hardware can be used to recognize the influence of certain environment classes on the human body.

When investigating self-assessed valence on five classes, on HR/IBI 44.31 % can be achieved, a class-reduction to three classes improves results to 59.48 %. Accuracy on a per class level decreases from good to bad consistently. This observation stays valid for GSR-data, while overall score rises to 65.53 % on five classes and 75.08 % on three classes. Feature fusion of HR/IBI and GSR improves results to 79.68% on five classes, where class *"4"* is a noticeable outlier at 53.76 %, more than 17 percent points lower than the next worst class *"5"*. Reduction from five to three classes leads to a more balanced model when it comes to per-class accuracy but only improving overall by less than two percent-points.

Both environment and user related models might be interpreted so that the cooling effect of the forest and the decrease in stress on the body is an important factor for identifying the surrounding via physiological data. Thus, results are promising, considering the distinction of different environments using physiological data, suggesting that it is possible to integrate the detection of an environment class (e.g., forest) with relaxing influence into wellbeing applications.

Considering practical use, the presented setup relied on consumer grade hardware only, which enables deployment of the shown approach to a wide audience.

7.5 Summary

This chapter extended SSI's field of operation towards recognizing wellbeing not just within the context of dyadic communication using social signals. The smart device in this chapter, forms a companion on the body, with a joined perspective rather than a conversational partner. Thus, not mimic or laughter, or even social cues are focus of the recognition process, but rather body reactions and environmental influences' characteristics.

Rapid prototyping was used for creating the record pipeline involving SSJ and MobileSSI and for the creation of mobile user interface to enable users to annotate "in the wild".

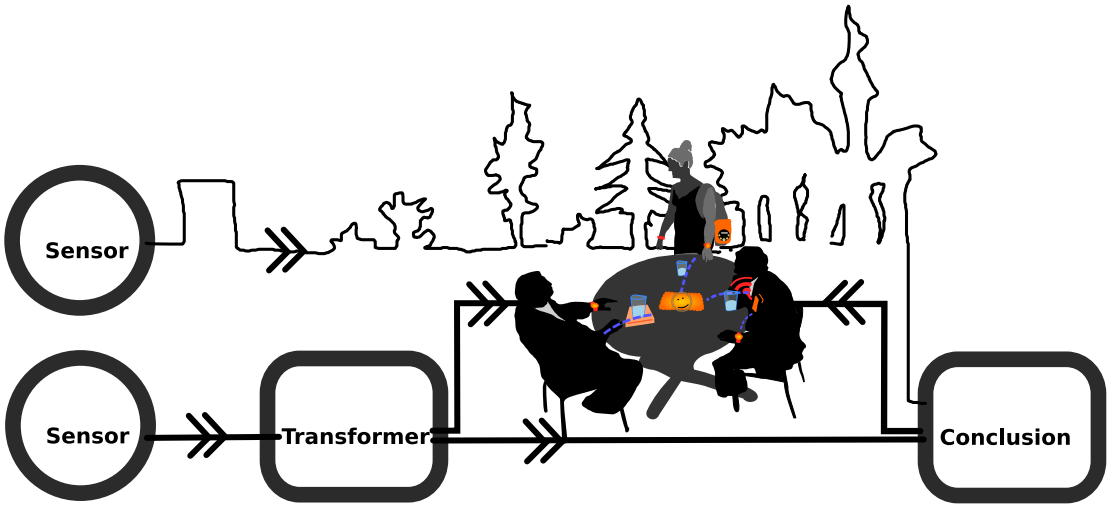
The influence of three environment classes (local climate zones), namely "dense trees" corresponding to forest, "scattered trees" (meadow) and "open midrise" (city) were studied. Moreover self assessed valence on five classes was also recorded using MobileSSI as another dependent variable. Recognition of valence and environment class was done based on audio on the one hand and physiological signals (GSR and BVP) on the other hand.

Feature level fusion of both physiological signals resulted in scores of 88.33 % on the three environment classes and 79.68 % on the five levels of self assessed valence.

Moreover audio based recognition was evaluated based on image features from DNNs (MobileNet V2) on spectrograms. Three environment classes could be recognized with 73.68% accuracy, whereas 61.79% were achieved with audio in the five-class valence problem.

MobileSSI enabled the combination of feature-sets created for in-the lab recordings , with consumer grade fitness-bands for unobstrusive use ”in the wild”. The processing and learning pipeline is light weight which enables it to run locally on mobile devices.

Chapter 8.
Conclusion



Today we are used to communication via smart devices to the extent, where we would feel the interaction with mobile devices comes naturally – we have to distinguish digital and analogue rather than natural and digital. With recent adoption of speech recognition, the transformation of input and interaction paradigms is still in progress. Devices such as smart watches become smaller and have to rely on active sensing, e.g. speech recognition. Thus, the initiative regarding input shifts from an active user to an active device.

Society is undergoing a similar transition, making use of mobile technology, as discourse on topics such as *M-Health* attests, e.g. in context of the COVID-19 crisis. To have an app that tracks human proximity and in that, the risk of infection, is seen as important part of overcoming the crisis. This discourse has also led to a decentralized approach with minimal data-

footprint. This shows both the topicality and importance of mobile processing of social signals under consideration of data privacy.

Mobile Social Signal Processing forms a community that mirrors that techno-social process within scientific research. Affective Computing and Social Signal Processing in the lab adapt to challenges ”in the wild”, where perspectives change from dyadic conversation, to accompaniment on the body.

The consideration of wellbeing as model of the users’ state, emerges from using models of emotion in human computer interaction, but focusing on longer periods of use as well as the necessity to rely more on environmental and situational context within mobile device usage. This thesis studied wellbeing social and emotional aspects regarding laughter recognition, in respect of drinking behavior and environment within local climate zones.

This work contributes the process of transforming SSI, a Social Signal Processing toolkit developed in the lab, into the wild, by porting it from Windows to Android. This required technological advancement, extending its areas of use to new circumstances of mobile devices and their usage in regard of wellbeing.

Laughter detection served as validation of the port, within the core domain of recognizing paralinguistic social cues. Drinking activity was chosen as use-case for interactive machine learning, to train machine learning models on users’ devices, with their help. Environmental context in relation to wellbeing was detected based on body worn sensors and labeling in the wild.

The studies presented in this thesis validate the technological advancement in hands on execution of new workflows.

8.1 Contributions

MobileSSI, presented in Chapter 4, contributes a software framework for mobile devices that enables a broad field of application. MobileSSI is expanding from the existing desktop implementation in aspects, such as rapid prototyping, coping with heterogeneous input, using fusion approaches, and integrated machine learning capabilities. It supports workflows of Social Signal Processing (SSP) such as recording, annotation, machine learning and real-time classification. Thus, it provides the basis of a tool to apply the methodology of SSP ”in the wild”, using mobile devices. MobileSSI is open-source, licensed under GPL and available online [1].

[1] https://github.com/hcmlab/SSI

Into the Wild

As a validation of the technical work (MobileSSI) and as an empirical realization employing it, a multi-modal laughter and group-enjoyment recognition, representing the core domain of SSP was conducted in Chapter 5. A social component was added to fusion using *multi-person fusion* and *rapidly prototyped* web based visualization. This addition was possible due to the high level of abstraction in the asynchronous fusion approach developed in the lab.

With newly integrated mobile capacities, recording "in the wild" was executed, to create a multi-modal corpus of six hours, involving three participants and two sessions. The created model lead to a live demonstration, presented at an international conference, combining two input devices with one displaying visualization of the recognized group enjoyment. The presented approach fits between fine-granular work on synchrony [214], not working in real time in contrast to the approach in MobileSSI and aggregation of individuals' affective states based on a Pleasure Arousal Dominance model [40], where interpretation of the outcome is not clear.

Interactive Machine Learning

Technical advancement and empirical validation can be found Chapter 6. Here the machine learning capabilities already integrated in SSI were extended with an *interactive machine learning* approach. Smart objects, in form of a digital drip mat, were used for labeling of drink behavior. Interactive machine learning conceptually is regarded as active and responsive, whereas user interaction is seen as defining characteristic. This on-device approach to machine learning, together with on-device real-time recognition tackle the challenge to respect users' *privacy* as presented in Chapter 6. An incremental (reactive) implementation of Naive Bayes is evaluated in an active learning scenario, querying the user based on a classification's certainty. The evaluation is based on a corpus containing 16.5 hours of data. Insights on interactive machine learning were generated by simulations based drink activity recorded using a smart watch. Here sampling based on high confidence intervals, is more effective with Naive Bayes whereas low confidence intervals are more efficient with linear SVMs. From recordings a base model was derived that was also foundation of users' bodystorming. While the prototype is perceived as intrusive in its requests for labels, users learned to simulate aspects of movement, that triggered recognition of a certain class. Drink activity recognition moves away from SSP's strict focus on social interaction but conveys the potential of reacting to situational context.

Wellbeing related Environmental Context

The core focus of Social Signal Processing is extended in Chapter 7, to also take *environmental context* into account as input, moving away from the picture of dyadic conversations that is underlying the study of laughter as social cue in Chapter 5. This change in concept is validated empirically with a field study. Physiological signals (GSR, BVP) are used to classify environments and users' wellbeing (valence) within selected local climate zones. MobileSSI was used for the recording of multi-modal data related to user, their self-assessment "in the wild" and their environment, resulting in a corpus of 26.6 hours of data by seven participants. *Rapid prototyping* was used to create annotation interfaces for a smart watch and a smart phone, that are suitable to labeling "in the wild". Classification of the local climate zone based on physiological data from consumer grade hardware, as well as audio data yielded similar results when relying on a single signal. Applying fusion on two physiological signals improved results considerably and underlined the importance of coping with *heterogeneous ubiquitous input*. This contribution adds objective measures of wellbeing in relation to environments and goes beyond applications based on user preferences of visual scenery as used e.g in routing applications [173].

Wellbeing, as presented in Chapter 2.4 is studied along four aspects over the course of this thesis. At first laughter is recognized from multi-modal data, representing *emotional* and *social* characteristic of wellbeing. Drink activity recognition is studied as aspect of *behavioral* wellbeing, where technology is enabled to interact with natural behavior to influence health related activities. Third *environmental* wellbeing is recognized based on local climate zones and their impact on physiological data.

Mobile Social Signal Processing enables M-Health technology to be more *predictive* by using classification, *personalized* by using machine learning on personal labels and *participatory* by using interactive machine learning.

8.2 Future Work

From this thesis' focus, the technological outlook can be found mainly in respect algorithmic evolvement and user integration within the process of machine learning in MSSP and M-Health.

Interactive machine learning [5] as well as federated machine learning [232] is a topic, that comes to mind naturally, when handling privacy-critical data. Instead of collecting data centrally, models are trained privately on the mobile devices the data originate from. Later on, the individual models are merged into one, that combines the decentrally acquired knowledge. Nonetheless, it can hardly be found in toolkits of Mobile Social Signal Processing or related topics, maybe since this approach contradicts the con-temporal principle of hoarding data, as a dragon would his treasure.

Interactive machine learning investigates a bottom up perspective to model generation, that focuses just on the user's private model, without giving an answer to the emergence of a unified knowledge pool. Future work would have to also explore solutions to combine individual models into one. Next to structural isolation by federation, there is a systemic approach to fairness [55] and privacy [54, 96] in machine learning. To reliably bind data-processing to ethical standards, those standards have to be implemented within software frameworks.

Motion, captured via accelerometers are a valuable, maybe the most valuable modality in mobile computing today, due to the wide availability. Therefore, interactive machine learning has been proposed for the design of movement interaction [80]. To enable the user not only to contribute to a machine learning model, but to also judge its quality, explainability might here meet embodiment [51] to give the user an intuitive grasp of an algorithm's doings.

Applying machine learning on a longer scope, with focus on predicting illness, MSSP can provide tools for preventive applications. Since MSSP involves models of Affective Computing, apps using MSSP can be aware of psycho-cognitve aspects and thus, contribute to the M-Health ecosystem [82].

THE END

Bibliography

[1] A. A. Adevi and F. Mårtensson, "Stress rehabilitation through garden therapy: The garden as a place in the recovery from stress," *Urban forestry & urban greening*, vol. 12, no. 2, pp. 230–237, 2013.

[2] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland, "Social fmri: Investigating and shaping social mechanisms in the real world," *Pervasive and Mobile Computing*, vol. 7, no. 6, pp. 643–659, 2011.

[3] U.-V. Albrecht, "Chancen und risiken von gesundheits-apps," in *Recht & Netz*. Nomos Verlagsgesellschaft mbH & Co. KG, 2018, pp. 417–430.

[4] T. M. Alschibaja, *Penisverletzungen bei Masturbation mit Staubsaugern*. Universitätsbibliothek München, 1978.

[5] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, December 2014.

[6] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. W. Schuller, "Snore sound classification using image-based deep spectrum features." in *INTERSPEECH*, 2017, pp. 3512–3516.

[7] J. Anderson, "Nudge: Improving decisions about health, wealth, and happiness, richard h. thaler and cass r. sunstein. yale university press, 2008. x+ 293 pages.[paperback edition, penguin, 2009, 320 pages.]," *Economics & Philosophy*, vol. 26, no. 3, pp. 369–376, 2010.

[8] E. André, J.-C. Martin, F. Lingenfelser, and J. Wagner, "Multimodal fusion in human-agent dialogue," *Coverbal Synchrony in Human-Machine Interaction*, pp. 387–410, 2014.

[9] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in neural information processing systems*, 2016, pp. 892–900.

[10] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[11] M. Bang, T. Timpka, H. Eriksson, E. Holm, C. Nordin *et al.*, "Mobile phone computing for in-situ cognitive behavioral therapy," in *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*. IOS Press, 2007, p. 1078.

[12] D. Bannach, O. Amft, and P. Lukowicz, "Rapid prototyping of activity recognition applications," *IEEE Pervasive Computing*, vol. 7, no. 2, pp. 22–31, 2008.

[13] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International conference on pervasive computing*, 2004, pp. 1–17.

[14] L. F. Barrett, "Valence is a basic building block of emotional life," *Journal of Research in Personality*, vol. 40, no. 1, pp. 35–55, 2006.

[15] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, p. 20–29, June 2004.

[16] T. Baur, I. Damian, F. Lingenfelser, J. Wagner, and E. André, "Nova: Automated analysis of nonverbal signals in social interactions," in *International Workshop on Human Behavior Understanding*. Springer, 2013, pp. 160–171.

[17] T. Baur, G. Mehlmann, I. Damian, F. Lingenfelser, J. Wagner, B. Lugrin, E. André, and P. Gebhard, "Context-aware automated analysis and annotation of social human-agent interactions," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 5, no. 2, p. 11, 2015.

[18] T. Beblo and S. Lautenbacher, *Neuropsychologie der Depression*. Hogrefe Verlag, 2006.

[19] C. Beck, A. Straub, S. Breitner, J. Cyrys, A. Philipp, J. Rathmann, A. Schneider, K. Wolf, and J. Jacobeit, "Air temperature characteristics of local climate zones in the augsburg urban area (bavaria, southern germany) under varying synoptic conditions," *Urban Climate*, vol. 25, pp. 152 – 166, 2018.

[20] V. Becker, L. Fessler, and G. Sörös, "Gestear: combining audio and motion sensing for gesture recognition on smartwatches," in *Proceedings of the 23rd International Symposium on Wearable Computers*, 2019, pp. 10–19.

[21] L. Berk, M. Prowse, G. Bains, J. Batt, J. Petrofsky, N. Daher, H. Danner, L. Ludeman, M. Lahman, S. Tan *et al.*, "Humor-associated laughter affects appetite hormones," *The FASEB Journal*, vol. 24, no. 1 Supplement, pp. 996–1, 2010.

[22] S. Bhattacharya and N. D. Lane, "From smart to deep: Robust activity recognition on smartwatches using deep learning," in *2016 IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom Workshops 2016, Sydney, Australia, March 14-18, 2016*, 2016, pp. 1–6.

[23] N. Biancone, C. Bicchielli, F. Ferri, and P. Grifoni, "Falls detection and assessment," in *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, New York, NY, USA, 2016, pp. 204–207.

[24] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "Moa: Massive online analysis," *J. Mach. Learn. Res.*, vol. 11, pp. 1601–1604, August 2010.

[25] B. Bittner, I. Aslan, C. T. Dang, and E. André, "Of smarthomes, iot plants, and implicit interaction design," in *Proceedings of the International Conference on Tangible, Embedded, and Embodied Interactions*, New York, NY, USA, 2019.

[26] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1. 05)[computer program]. retrieved may 1, 2009," 2009.

[27] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [computer program]. version 6.0. 37," *RetrievedFebruary*, vol. 3, p. 2018, 2018.

[28] R. A. Bolt, ""put-that-there" voice and gesture at the graphics interface," in *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, 1980, pp. 262–270.

[29] R. Brueckner and B. Schuller, "Social Signal Classification using Deep BLSTM Recurrent Neural Networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4823–4827.

[30] M. Burke, C. Marlow, and T. Lento, "Social network activity and social well-being," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2010, pp. 1909–1912.

[31] T. K. Burki, "Modern art stinks," *The Lancet Respiratory Medicine*, vol. 6, no. 6, p. 416, 2018.

[32] A. Capponi, C. Fiandrino, B. Kantarci, L. Foschini, D. Kliazovich, and P. Bouvry, "A survey on mobile crowdsensing systems: Challenges, solutions, and opportunities," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2419–2465, thirdquarter 2019.

[33] I. Carreras, A. Matic, P. Saar, and V. Osmani, "Comm2sense: Detecting proximity through smartphones," in *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, 2012, pp. 253–258.

[34] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani, "Dynamics of person-to-person interactions from distributed rfid sensor networks," *PloS one*, vol. 5, no. 7, 2010.

[35] Y.-P. Chen, J.-Y. Yang, S.-N. Liou, G.-Y. Lee, and J.-S. Wang, "Online classifier construction algorithm for human activity detection using a tri-axial accelerometer," *Applied Mathematics and Computation*, vol. 205, pp. 849–860, 2008.

[36] P. Cohen, "Foundations of collaborative task-oriented dialogue: What's in a slot?" in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Stockholm, Sweden, September 2019, pp. 198–209.

[37] M. Conti, S. Giordano, M. May, and A. Passarella, "From opportunistic networks to opportunistic computing," *IEEE Communications Magazine*, vol. 48, no. 9, pp. 126–139, Sep. 2010.

[38] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[39] S. Cosentino, S. Sessa, and A. Takanishi, "Quantitative laughter detection, measurement, and classification—a critical survey," *IEEE Reviews in Biomedical Engineering*, vol. 9, pp. 148–162, 2016.

[40] C. Coutrix, G. Jacucci, I. Advouevski, V. Vervondel, M. Cavazza, S. W. Gilroy, and L. Parisi, "Supporting multi-user participation with affective multimodal fusion," in *2011 Ninth International Conference on Creating, Connecting and Collaborating through Computing*. IEEE, 2011, pp. 24–31.

[41] S. Cristina and K. P. Camilleri, "Unobtrusive and pervasive video-based eye-gaze tracking," *Image and Vision Computing*, vol. 74, pp. 21–40, 2018.

[42] A. Crooks, A. Croitoru, A. Stefanidis, and J. Radzikowski, "# earthquake: Twitter as a distributed sensor system," *Transactions in GIS*, vol. 17, no. 1, pp. 124–147, 2013.

[43] N. Cummins, B. Vlasenko, H. Sagha, and B. Schuller, "Enhancing speech-based depression detection through gender dependent vowel-level formant features," in *Conference on Artificial Intelligence in Medicine in Europe*. Springer, 2017, pp. 209–214.

[44] I. Damian, M. Dietz, and E. André, "The ssj framework: Augmenting social interactions using mobile signal processing and live feedback," *Frontiers in ICT*, vol. 5, p. 13, 2018.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[46] L. Dennison, L. Morrison, G. Conway, and L. Yardley, "Opportunities and challenges for smartphone applications in supporting health behavior change: Qualitative study," *J Med Internet Res*, vol. 15, no. 4, p. e86, Apr 2013.

[47] D. Destoumieux-Garzón, P. Mavingui, G. Boetsch, J. Boissier, F. Darriet, P. Duboz, C. Fritsch, P. Giraudoux, F. Le Roux, S. Morand, C. Paillard, D. Pontier, C. Sueur, and Y. Voituron, "The one health concept: 10 years old and a long road ahead," *Frontiers in Veterinary Science*, vol. 5, p. 14, 2018.

[48] E. Di Lascio, S. Gashi, and S. Santini, "Laughter recognition using non-invasive wearable devices," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, New York, NY, USA, 2019, p. 262–271.

[49] M. Dietz, D. Schork, I. Damian, A. Steinert, M. Haesner, and E. André, "Automatic detection of visual search for the elderly using eye and head tracking data," *KI-Künstliche Intelligenz*, vol. 31, no. 4, pp. 339–348, 2017.

[50] M. Dietz, I. Aslan, D. Schiller, S. Flutura, A. Steinert, R. Klebbe, and E. André, "Stress annotations from older adults - exploring the foundations for mobile ml-based health assistance," in *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare*, New York, NY, USA, 2019, pp. 149–158.

[51] P. Dourish, *Where the action is: the foundations of embodied interaction*. MIT press, 2004.

[52] M. Dunlop and S. Brewster, "The challenge of mobile devices for human computer interaction," *Personal and ubiquitous computing*, vol. 6, no. 4, pp. 235–236, 2002.

[53] S. Dupont and J. Luettin, "Audio-visual Speech Modeling for Continuous Speech Recognition," *Transactions on Multimedia*, vol. 2(3), pp. 141 – 151, 2000.

[54] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[55] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," *arXiv preprint arXiv:1104.3913*, 2011.

[56] M. Eid and R. J. Larsen, *The science of subjective well-being*. Guilford Press, 2008.

[57] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16, 1999.

[58] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[59] A. A. El-Hilly, S. S. Iqbal, M. Ahmed, Y. Sherwani, M. Muntasir, S. Siddiqui, Z. Al-Fagih, O. Usmani, and A. B. Eisingerich, "Game on? smoking cessation through the gamification of mhealth: A longitudinal qualitative study," *JMIR serious games*, vol. 4, no. 2, p. e18, 2016.

[60] B. Endrass, I. Damian, P. Huber, M. Rehm, and E. André, "Generating culture-specific gestures for virtual agent dialogs," in *Intelligent Virtual Agents*, Berlin, Heidelberg, 2010, pp. 329–335.

[61] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. Al'Absi, and S. Shah, "Autosense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field," in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2011, pp. 274–287.

[62] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April 2016.

[63] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor," in *21st ACM International Conference on Multimedia*, 2013, pp. 835–838.

[64] J. A. Fails and D. R. Olsen, Jr., "Interactive machine learning," in *Proceedings of the 8th International Conference on Intelligent User Interfaces*, New York, NY, USA, 2003, pp. 39–45.

[65] S. Flutura, J. Wagner, F. Lingenfelser, A. Seiderer, and E. André, "Mobilessi - a multi-modal framework for social signal interpretation on mobile devices," in *2016 12th International Conference on Intelligent Environments (IE)*, 2016, pp. 210–213.

[66] S. Flutura, J. Wagner, F. Lingenfelser, A. Seiderer, and E. André, "MobileSSI: Asynchronous fusion for social signal interpretation in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2016, pp. 266–273.

[67] S. Flutura, J. Wagner, F. Lingenfelser, A. Seiderer, and E. André, "Laughter detection in the wild: Demonstrating a tool for mobile social signal processing and visualization," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2016, p. 406–407.

[68] S. Flutura, A. Seiderer, I. Aslan, C. T. Dang, R. Schwarz, D. Schiller, and E. André, "DrinkWatch: A mobile wellbeing application based on interactive and cooperative machine learning," in *Proceedings of the International Conference on Digital Health*, 2018, pp. 65–74.

[69] S. Flutura, A. Seiderer, I. Aslan, M. Dietz, D. Schiller, C. Beck, J. Rathmann, and E. André, "Mobile sensing for wellbeing estimation of urban green using physiological signals," in *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good*, New York, NY, USA, 2019, p. 249–254.

[70] S. Flutura, A. Seiderer, T. Huber, K. Weitz, I. Aslan, R. Schlagowski, M. Dietz, J. Rathmann, and E. André, "Interactive machine learning and explainability in mobile classification of forest-aesthetics," in *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good*, New York, NY, USA, 2020, p. 6.

[71] B. J. Fogg, "Persuasive technology: using computers to change what we think and do," *Ubiquity*, vol. 2002, no. December, p. 5, 2002.

[72] B. H. L. France, A. D. Heisel, and M. J. Beatty, "Is there empirical evidence for a nonverbal profile of extraversion?: a meta-analysis and critique of the literature," *Communication Monographs*, vol. 71, no. 1, pp. 28–48, 2004.

[73] C. E. Frank, D. Naugler, and M. Traina, "Teaching user interface prototyping," *Journal of Computing Sciences in Colleges*, vol. 20, no. 6, pp. 66–73, 2005.

[74] L. Frank, J. F. Sallis, T. L. Conway, J. Chapman, B. Saelens, and W. Bachman, "Many pathways from land use to health: Associations between neighborhood walkability and active transportation, body mass index, and air quality," *Journal of the American Planning Association*, vol. 72, pp. 75–87, 03 2006.

[75] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[76] K. Fukumoto, T. Terada, and M. Tsukamoto, "A smile/laughter recognition mechanism for smile-based life logging," in *Proc. of Augmented Human International Conference*, 2013, pp. 213–220.

[77] J. Gama, "Iterative naive bayes," in *Discovery Science*, Berlin, Heidelberg, 1999, pp. 80–91.

[78] W. M. Gesler, "Therapeutic landscapes: Theory and a case study of epidauros, greece," *Environment and Planning D: Society and Space*, vol. 11, no. 2, pp. 171–189, 1993.

[79] A. Ghosh and G. Riccardi, "Recognizing human activities from smartphone sensor signals," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, New York, NY, USA, 2014, pp. 865–868.

[80] M. Gillies, "Understanding the role of interactive machine learning in movement interaction design," *ACM Trans. Comput.-Hum. Interact.*, vol. 26, no. 1, February 2019.

[81] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Processing*, vol. 150, 01 2009.

[82] A. Gorini, K. Mazzocco, S. Triberti, V. Sebri, L. Savioni, and G. Pravettoni, "A p5 approach to m-health: Design suggestions for advanced mobile health technology," *Frontiers in Psychology*, vol. 9, p. 2066, 2018.

[83] B. Guo, Z. Yu, X. Zhou, and D. Zhang, "From participatory sensing to mobile crowd sensing," in *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*. IEEE, 2014, pp. 593–598.

[84] B. Guo, Z. Wang, Z. Yu, Y. Wang, N. Y. Yen, R. Huang, and X. Zhou, "Mobile crowd sensing and computing: The review of an emerging human-ed sensing paradigm," *ACM Comput. Surv.*, vol. 48, no. 1, pp. 7:1–7:31, August 2015.

[85] G. Hagerer, N. Cummins, F. Eyben, and B. W. Schuller, "" did you laugh enough today?"-deep neural networks for mobile and wearable laughter trackers." in *INTERSPEECH*, 2017, pp. 2044–2045.

[86] G. Hagerer, N. Cummins, F. Eyben, and B. W. Schuller, "Robust laughter detection for wearable wellbeing sensing," in *Proceedings of the 2018 International Conference on Digital Health, DH 2018, Lyon, France, April 23-26, 2018*, 2018, pp. 156–157.

[87] D. Hasenfratz, O. Saukh, S. Sturzenegger, and L. Thiele, "Participatory air pollution monitoring using smartphones," *Mobile Sensing*, vol. 1, pp. 1–5, 2012.

[88] J. F. Helliwell and R. D. Putnam, "The social context of well–being," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 359, no. 1449, pp. 1435–1446, 2004.

[89] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[90] H. Hsu, K. Tsai, Z. Cheng, and T. Huang, "Posture recognition with g-sensors on smart phones," in *2012 15th International Conference on Network-Based Information Systems*, Sep. 2012, pp. 588–591.

[91] Z. Huang, J. Epps, D. Joachim, and M. Chen, "Depression detection from short utterances via diverse smartphones in natural environmental conditions." in *Interspeech*, 2018, pp. 3393–3397.

[92] E. Hutchins, *Cognition in the Wild*, no. 1995. MIT press, 1995.

[93] T. Huynh and B. Schiele, "Analyzing features for activity recognition," in *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies*, New York, NY, USA, 2005, pp. 159–163.

[94] W. James, "The principles of psychology, vol. 2. ny, us: Henry holt and company," 1890.

[95] M.-p. Jansen, D. K. Heylen, K. P. Truong, G. Englebienne, and D. S. Nazareth, "The mulai corpus: Multimodal recordings of spontaneous laughter in dyadic interaction," in *Proceedings of Laughter Workshop*, 2018, pp. 58–63.

[96] K. Jarmul, "Privacy: the last stand for fair algorithms," 10 2018, talk at the Strange Loop Conference.

[97] Z. Jianyong, L. Haiyong, C. Zili, and L. Zhaohui, "Rssi based bluetooth low energy indoor positioning," in *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Oct 2014, pp. 526–533.

[98] K. Kalimeri and C. Saitis, "Exploring multimodal biosignal features for stress detection during indoor mobility," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2016, pp. 53–60.

[99] A. Kapadia, N. Triandopoulos, C. Cornelius, D. Peebles, and D. Kotz, "Anonysense: Opportunistic and privacy-preserving context collection," in *International Conference on Pervasive Computing*, 05 2008, pp. 280–297.

[100] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and Biological Engineering and Computing*, vol. 42, no. 3, pp. 419–427, May 2004.

[101] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[102] D. E. Knuth, *The Art of Computer Programming, Volume 2 (2nd Ed.): Seminumerical Algorithms*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1985.

[103] J. Kukkonen, E. Lagerspetz, P. Nurmi, and M. Andersson, "Betelgeuse: A platform for gathering and processing situational data," *IEEE Pervasive Computing*, vol. 8, no. 2, pp. 49–56, 2009.

[104] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[105] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[106] O. Lara and M. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, pp. 1192–1209, 01 2013.

[107] Y. Lee and M. Song, "Using a smartwatch to detect stereotyped movements in children with developmental disabilities," *IEEE Access*, vol. 5, pp. 5506–5514, 2017.

[108] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 1994, pp. 3–12.

[109] F. Lingenfelser, J. Wagner, and E. André, "A systematic discussion of fusion techniques for multi-modal affect recognition tasks," in *13th International Conference on Multimodal Interfaces*, 2011, pp. 19–26.

[110] F. Lingenfelser, J. Wagner, E. André, G. McKeown, and W. Curran, "An Event Driven Fusion Approach for Enjoyment Recognition in Real-time," in *22nd ACM International Conference on Multimedia*, 2014, pp. 377–386.

[111] J. Liono, T. Nguyen, P. P. Jayaraman, and F. D. Salim, "Ute: A ubiquitous data exploration platform for mobile sensing experiments," in *2016 17th IEEE International Conference on Mobile Data Management (MDM)*, June 2016, pp. 349–352.

[112] H. Lu, W. Pan, N. D. Lane, T. Choudhury, and A. T. Campbell, "Soundsense: Scalable sound sensing for people-centric applications on mobile phones," in *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*, New York, NY, USA, 2009, pp. 165–178.

[113] H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell, "The jigsaw continuous sensing engine for mobile phone applications," in *Proceedings of the 8th ACM conference on embedded networked sensor systems*. ACM, 2010, pp. 71–84.

[114] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," *CoRR*, vol. abs/1906.08172, 2019.

[115] M. Mancini, G. Varni, D. Glowinski, and G. Volpe, "Computing and evaluating the body laughter index," in *Human Behavior Understanding*. Springer, 2012, vol. 7559, pp. 90–98.

[116] M. Mancini, L. Ach, E. Bantegnie, T. Baur, N. Berthouze, D. Datta, Y. Ding, S. Dupont, H. J. Griffin, F. Lingenfelser, R. Niewiadomski, C. Pelachaud, O. Pietquin, B. Piot, J. Urbain, G. Volpe, and J. Wagner, "Laugh when you're winning," in *Innovative and Creative Developments in Multimodal Interaction Systems*. Springer, 2014, vol. 425, pp. 50–79.

[117] E. Marchi, F. Eyben, G. Hagerer, and B. W. Schuller, "Real-time tracking of speakers' emotions, states, and traits on mobile platforms." in *INTERSPEECH*, 2016, pp. 1182–1183.

[118] S. Marsella, J. Gratch, P. Petta *et al.*, "Computational models of emotion," *A Blueprint for Affective Computing-A sourcebook and manual*, vol. 11, no. 1, pp. 21–46, 2010.

[119] W. M. Marston, *Emotions of normal people*. Kegan Paul Trench Trubner And Company., Limited, 1928.

[120] T. M. Marteau, D. Ogilvie, M. Roland, M. Suhrcke, and M. P. Kelly, "Judging nudging: can nudging improve population health?" *Bmj*, vol. 342, p. d228, 2011.

[121] A. Masciadri, M. Sacchi, S. Comai, and F. Salice, "Wellness indexes to assess quality of life: A technological support," in *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good*, New York, NY, USA, 2019, pp. 213–218.

[122] S. McCallum, "Gamification and serious games for personalized health." in *pHealth*, 2012, pp. 85–96.

[123] G. R. McGregor and J. K. Vanos, "Heat: a primer for public health researchers," *Public health*, vol. 161, pp. 138–146, 2018.

[124] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, p. 746, 1976.

[125] G. McKeown, W. Curran, C. McLoughlin, H. J. Griffin, and N. Bianchi-Berthouze, "Laughter induction techniques suitable for generating motion capture data of laughter associated body movements," in *Proc. of 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–5.

[126] G. McKeown, W. Curran, J. Wagner, F. Lingenfelser, and E. André, "The belfast storytelling database – a spontaneous social interaction database with laughter focused annotation," in *International Conference on Affective Computing and Intelligent Interaction*, 2015.

[127] P. Melville and R. J. Mooney, "Diverse ensembles for active learning," in *Proceedings of the Twenty-First International Conference on Machine Learning*, New York, NY, USA, 2004, p. 74.

[128] F. Metze, S. Rawat, and Y. Wang, "Improved audio features for large-scale multimedia event detection," in *2014 IEEE International Conference on Multimedia and Expo (ICME)*, July 2014, pp. 1–6.

[129] S. L. Middendorf, "Autarkie als (selbst-) reβ // this should be curled beta,lexion," *Vielfalt und Veränderung*, p. 51, 2017.

[130] E. Miluzzo, N. D. Lane, K. Fodor, R. Peterson, H. Lu, M. Musolesi, S. B. Eisenman, X. Zheng, and A. T. Campbell, "Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application," in *Proceedings of the 6th ACM conference on Embedded network sensor systems.* ACM, 2008, pp. 337–350.

[131] T. Miu, P. Missier, and T. Plötz, "Bootstrapping personalised human activity recognition models using online active learning," in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, Oct 2015, pp. 1138–1147.

[132] D. A. Moses, N. Mesgarani, M. K. Leonard, and E. F. Chang, "Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity," *Journal of Neural Engineering*, vol. 13, no. 5, p. 056004, aug 2016.

[133] M. Nakayoshi, M. Kanda, R. Shi, and R. de Dear, "Outdoor thermal physiology along human pathways: a study using a wearable measurement system," *International Journal of Biometeorology*, vol. 59, no. 5, pp. 503–515, 04 2015.

[134] Y. Nam, S. Rho, and C. Lee, "Physical activity recognition using multiple sensors embedded in a wearable device," *ACM Trans. Embed. Comput. Syst.*, vol. 12, no. 2, pp. 26:1–26:14, February 2013.

[135] D. Navarre, P. Palanque, R. Bastide, A. Schyn, M. Winckler, L. P. Nedel, and C. M. D. S. Freitas, "A Formal Description of Multimodal Interaction Techniques for Immersive Virtual Reality Applications," in *IFIP TC13 International Conference on Human-Computer Interaction (INTERACT)*, 2005, pp. 170–183.

[136] S. Nirjon, R. F. Dickerson, P. Asare, Q. Li, D. Hong, J. A. Stankovic, P. Hu, G. Shen, and X. Jiang, "Auditeur: A mobile-cloud service platform for acoustic event detection on smartphones," in *Proceeding of the 11th annual international conference on Mobile systems, applications, and services.* ACM, 2013, pp. 403–416.

[137] F. Okeke, M. Sobolev, N. Dell, and D. Estrin, "Good vibrations: Can a digital nudge reduce digital overload?" in *Proceedings of the 20th International Conference on*

Human-Computer Interaction with Mobile Devices and Services, New York, NY, USA, 2018, pp. 4:1–4:12.

[138] S. Oniani, I. M. Pires, N. M. Garcia, I. Mosashvili, and N. Pombo, "A review of frameworks on continuous data acquisition for e-health and m-health," in *Proceedings of the 5th EAI International Conference on Smart Objects and Technologies for Social Good*. ACM, 2019, pp. 231–234.

[139] W. H. Organization *et al.*, "Constitution of the world health organization," 1995.

[140] A. Ortony and T. J. Turner, "What's basic about basic emotions?" *Psychological review*, vol. 97, no. 3, p. 315, 1990.

[141] A. Oulasvirta, E. Kurvinen, and T. Kankainen, "Understanding contexts by being there: case studies in bodystorming," *Personal and ubiquitous computing*, vol. 7, no. 2, pp. 125–134, 2003.

[142] S. Oviatt, "Designing robust multimodal systems for universal access," in *Proceedings of the 2001 EC/NSF Workshop on Universal Accessibility of Ubiquitous Computing: Providing for the Elderly*, New York, NY, USA, 2001, p. 71–74.

[143] N. Palaghias, S. A. Hoseinitabatabaei, M. Nati, A. Gluhak, and K. Moessner, "A survey on mobile social signal processing," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, p. 57, 2016.

[144] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[145] B. J. Park, Y. Tsunetsugu, T. Kasetani, T. Kagawa, and Y. Miyazaki, "The physiological effects of shinrin-yoku (taking in the forest atmosphere or forest bathing): evidence from field experiments in 24 forests across japan," *Environmental health and preventive medicine*, vol. 15, no. 1, p. 18, 2010.

[146] R. Parsons and T. C. Daniel, "Good looking: in defense of scenic landscape aesthetics," *Landscape and Urban Planning*, vol. 60, no. 1, pp. 43 – 56, 2002.

[147] A. Pentland, "Social signal processing [exploratory dsp]," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 108–111, July 2007.

[148] R. W. Picard, *Affective computing*. MIT press, 2000.

[149] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[150] M. Pielot, T. Dingler, J. S. Pedro, and N. Oliver, "When attention is not scarce - detecting boredom from mobile phone usage," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, New York, NY, USA, 2015, p. 825–836.

[151] S. Pittner and S. V. Kamarthi, "Feature extraction from wavelet coefficients for pattern recognition tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 1, pp. 83–88, 1999.

[152] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[153] R. A. Plunz, Y. Zhou, M. I. C. Vintimilla, K. Mckeown, T. Yu, L. Uguccioni, and M. P. Sutto, "Twitter sentiment in new york city parks as measure of well-being," *Landscape and urban planning*, vol. 189, pp. 235–246, 2019.

[154] R. Plutchik, "A psychoevolutionary theory of emotions," 1982.

[155] E. Politou, E. Alepis, and C. Patsakis, "A survey on mobile affective computing," *Computer Science Review*, vol. 100, no. 25, pp. 79–100, 2017.

[156] A. Polychroniou, "The sspnet-mobile corpus: from the detection of non-verbal cues to the inference of social behaviour during mobile phone conversations," Ph.D. dissertation, University of Glasgow, 2014.

[157] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, "Emotionsense: a mobile phones based adaptive platform for experimental social psychology research," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 2010, pp. 281–290.

[158] K. K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow, "Sociablesense: exploring the trade-offs of adaptive sampling and computation offloading for social sensing," in *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 2011, pp. 73–84.

[159] R. K. Rana, C. T. Chou, S. S. Kanhere, N. Bulusu, and W. Hu, "Ear-phone: An end-to-end participatory urban noise mapping system," in *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, New York, NY, USA, 2010, pp. 105–116.

[160] J. Rathmann and S. Brumann, "Therapeutische landschaften in der psychoonkologie." *Gaia: Okologische Perspektiven in Natur-, Geistes- und Wirtschaftswissenschaften*, vol. 26, pp. 254–258, 01 2017.

[161] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence - Volume 3*, 2005, pp. 1541–1546.

[162] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in *Aaai*, 2005, pp. 1541–1546.

[163] R. Rawassizadeh, B. A. Price, and M. Petre, "Wearables: Has the age of smartwatches finally arrived?" *Commun. ACM*, vol. 58, no. 1, pp. 45–47, December 2014.

[164] P. Renaud and J.-P. Blondin, "The stress of stroop performance: Physiological and emotional responses to color–word interference, task pacing, and pacing speed," *International Journal of Psychophysiology*, vol. 27, no. 2, pp. 87–97, 1997.

[165] G. Revill, "How is space made in sound? spatial mediation, critical phenomenology and the political agency of sound," *Progress in Human Geography*, vol. 40, no. 2, pp. 240–256, 2016.

[166] V. Rideout, S. Fox *et al.*, "Digital health practices, social media use, and mental well-being among teens and young adults in the us," 2018.

[167] H. Ritschel, A. Seiderer, K. Janowski, I. Aslan, and E. André, "Drink-o-mender: An adaptive robotic drink adviser," in *Proceedings of the 3rd International Workshop on Multisensory Approaches to Human-Food Interaction.* ACM, 2018, p. 3.

[168] N. D. Rodríguez, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores, "A survey on ontologies for human behavior recognition," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 43:1–43:33, March 2014.

[169] Y. Rogers, "Interaction design gone wild: Striving for wild theory," *Interactions*, vol. 18, no. 4, pp. 58–62, July 2011.

[170] M. Rossi, O. Amft, S. Feese, C. Käslin, and G. Tröster, "Myconverse in action: monitoring conversations using smartphones." in *UbiComp (Adjunct Publication)*, 2013, pp. 1307–1308.

[171] M. Rossi, S. Feese, O. Amft, N. Braune, S. Martis, and G. Tröster, "Ambientsense: A real-time ambient sound recognition system for smartphones," in *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops).* IEEE, 2013, pp. 230–235.

[172] D. Rubine, "Combining gestures and direct manipulation," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1992, pp. 659–660.

[173] N. Runge, P. Samsonov, D. Degraen, and J. Schöning, "No more autobahn!: Scenic route generation using googles street view," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, New York, NY, USA, 2016, pp. 147–151.

[174] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[175] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[176] V. Sacharin, K. Schlegel, and K. R. Scherer, "Geneva emotion wheel rating study," Center for Person Kommunikation, Aalborg University NCCR Affective Sciences, Copenhagen, Denmark, Technical Report, 2012.

[177] K. M. Sagayam and D. J. Hemanth, "Hand posture and gesture recognition techniques for virtual reality applications: a survey," *Virtual Reality*, vol. 21, no. 2, pp. 91–107, 2017.

[178] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[179] P. Samsonov, F. Heller, and J. Schöning, "Autobus: Selection of passenger seats based on viewing experience for touristic tours," in *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia*, New York, NY, USA, 2017, pp. 321–326.

[180] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[181] A. Sano, P. Johns, and M. Czerwinski, "Designing opportune stress intervention delivery timing using multi-modal data," in *ACII 2017*, October 2017.

[182] D. Schiller, K. Weitz, K. Janowski, and E. André, "Human-inspired socially-aware interfaces," in *International Conference on Theory and Practice of Natural Computing*. Springer, 2019, pp. 41–53.

[183] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous *et al.*, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[184] A. J. Schwartz, P. S. Dodds, J. P. O'Neil-Dunne, C. M. Danforth, and T. H. Ricketts, "Visitors to urban greenspace have higher sentiment and lower negativity on twitter," *People and Nature*, vol. 1, no. 4, pp. 476–485, 2019.

[185] A. Seiderer and E. André, "Development of a multi-device nutrition logging prototype including a smartscale," in *Proceedings of the 2017 International Conference on Digital Health*, New York, NY, USA, 2017, pp. 239–240.

[186] A. Seiderer, S. Flutura, and E. André, "Development of a mobile multi-device nutrition logger," in *Proceedings of the 2Nd ACM SIGCHI International Workshop on Multisensory Approaches to Human-Food Interaction*, New York, NY, USA, 2017, pp. 5–12.

[187] A. Seiderer, S. Flutura, and E. André, "Development of a mobile multi-device nutrition logger," in *Proceedings of the 2nd ACM SIGCHI International Workshop on Multisensory Approaches to Human-Food Interaction*, New York, NY, USA, 2017, p. 5–12.

[188] A. Seiderer, M. Dietz, I. Aslan, and E. André, "Enabling privacy with transfer learning for image classification dnns on mobile devices," in *Proceedings International Conference on Smart Objects and Technologies for Social Good*, New York, NY, USA, 2018, pp. 25–30.

[189] M. E. Seligman, *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment.* Simon and Schuster, 2004.

[190] P. Sengers, K. Boehner, S. David, and J. J. Kaye, "Reflective design," in *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*, New York, NY, USA, 2005, pp. 49–58.

[191] D. R. Seshadri, E. V. Davies, E. R. Harlow, J. J. Hsu, S. C. Knighton, T. A. Walker, J. E. Voos, and C. K. Drummond, "Wearable sensors for covid-19: A call to action to harness our digital infrastructure for remote patient monitoring and virtual assessments," *Frontiers in Digital Health*, vol. 2, p. 8, 2020.

[192] B. Settles, "Active learning literature survey," *Computer Sciences Technical Report*, vol. 1648, 2010.

[193] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, New York, NY, USA, 1992, pp. 287–294.

[194] F. Shahmohammadi, A. Hosseini, C. E. King, and M. Sarrafzadeh, "Smartwatch based activity recognition using active learning," in *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, July 2017, pp. 321–329.

[195] F. Shahmohammadi, A. Hosseini, C. E. King, and M. Sarrafzadeh, "Smartwatch based activity recognition using active learning," in *Proceedings of the Second IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies*, 2017, p. 321–329.

[196] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: primal estimated sub-gradient solver for svm," *Mathematical Programming*, vol. 127, no. 1, pp. 3–30, Mar 2011.

[197] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. M. Havinga, "Complex human activity recognition using smartphone and wrist-worn motion sensors," *Sensors*, vol. 16, no. 4, 2016.

[198] J. Shu, M. Chiu, and P. Hui, "Emotion sensing for mobile computing," *IEEE Communications Magazine*, vol. 57, no. 11, pp. 84–90, November 2019.

[199] A. Singh, N. Bianchi-Berthouze, and A. C. Williams, "Supporting everyday function in chronic pain using wearable technology," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2017, p. 3903–3915.

[200] C. Song, H. Ikei, M. Igarashi, M. Miwa, M. Takagaki, and Y. Miyazaki, "Physiological and psychological responses of young males during spring-time walks in urban parks," *Journal of Physiological Anthropology*, vol. 33, no. 1, p. 8, 04 2014.

[201] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, Aug 2000.

[202] I. D. Stewart and T. R. Oke, "Local climate zones for urban temperature studies," *Bulletin of the American Meteorological Society*, vol. 93, no. 12, pp. 1879–1900, 2012.

[203] C. Strohrmann, R. Labruyère, C. N. Gerber, H. J. van Hedel, B. Arnrich, and G. Tröster, "Monitoring motor capacity changes of children during rehabilitation using body-worn sensors," *Journal of NeuroEngineering and Rehabilitation*, vol. 10, no. 1, p. 83, Jul 2013.

[204] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, "Activity-aware mental stress detection using physiological sensors," in *International conference on Mobile computing, applications, and services*. Springer, 2010, pp. 282–301.

[205] N. A. Syed, S. Huan, L. Kah, and K. Sung, "Incremental learning with support vector machines," 1999.

[206] L. Tan, K. Zhang, K. Wang, X. Zeng, X. Peng, and Y. Qiao, "Group emotion recognition with individual facial emotion cnns and global image based cnns," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 549–552.

[207] K. Taraldsen, S. F. Chastin, I. I. Riphagen, B. Vereijken, and J. L. Helbostad, "Physical activity monitoring by use of accelerometer-based body-worn sensors in older adults: A

systematic literature review of current knowledge and applications," *Maturitas*, vol. 71, no. 19, 2017.

[208] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, March 2002.

[209] M. Tonsen, J. Steil, Y. Sugano, and A. Bulling, "Invisibleeye: Mobile eye tracking using multiple low-resolution cameras and learning-based gaze estimation," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 106:1–106:21, September 2017.

[210] C.-H. Tsai, C.-Y. Lin, and C.-J. Lin, "Incremental and decremental training for linear classification," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2014, pp. 343–352.

[211] H. Tsujita and J. Rekimoto, "Smile-encouraging digital appliances," *IEEE Pervasive Computing*, vol. 12, no. 4, pp. 5–7, Oct 2013.

[212] R. S. Ulrich, "View through a window may influence recovery from surgery," *Science*, vol. 224, no. 4647, pp. 420–421, 1984.

[213] Y. Vaizman, K. Ellis, G. Lanckriet, and N. Weibel, "Extrasensory app: Data collection in-the-wild with rich user interface to self-report behavior," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2018, pp. 554:1–554:12.

[214] G. Varni, M. Avril, A. Usta, and M. Chetouani, "Syncpy: A unified open-source analytic library for synchrony," in *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*, New York, NY, USA, 2015, pp. 41–47.

[215] E. Velten Jr, "A laboratory task for induction of mood states," *Behaviour research and therapy*, vol. 6, no. 4, pp. 473–482, 1968.

[216] A. Vinciarelli and A. S. Pentland, "New social signals in a new interaction world: The next frontier for social signal processing," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 1, no. 2, pp. 10–17, April 2015.

[217] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743 – 1759, 2009, visual and multimodal analysis of human spontaneous behaviour:.

[218] A. Vinciarelli, R. Murray-Smith, and H. Bourlard, "Mobile social signal processing: vision and research issues," in *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services.* ACM, 2010, pp. 513–516.

[219] T. Vogt, E. André, and N. Bee, "EmoVoice - a framework for online recognition of emotions from voice," in *Perception in Multimodal Dialogue Systems*. Springer, 2008, vol. 5078, pp. 188–199.

[220] J. Wagner and E. André, "Real-time sensing of affect and social signals in a multimodal framework: a practical approach," in *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*, 2018, pp. 227–261.

[221] J. Wagner, F. Lingenfelser, E. André, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 206–218, 2011.

[222] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The Social Signal Interpretation (SSI) framework: Multimodal signal processing and recognition in real-time," in *21st ACM International Conference on Multimedia*, 2013, pp. 831–834.

[223] J. Wagner, D. Schiller, A. Seiderer, and E. André, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?" in *Proc. Interspeech 2018*, 2018, pp. 147–151.

[224] Y. Wang, J. Lin, M. Annavaram, Q. A. Jacobson, J. Hong, B. Krishnamachari, and N. Sadeh, "A framework of energy efficient mobile sensing for automatic user state recognition," in *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*, New York, NY, USA, 2009, pp. 179–192.

[225] M. Weiser, "The computer for the 21st century," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 3, no. 3, pp. 3–11, July 1999.

[226] G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber, "Smartwatch-based activity recognition: A machine learning approach," in *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2016, pp. 426–429.

[227] K. Weitz, T. Hassan, U. Schmid, and J. Garbas, "Towards explaining deep learning networks to distinguish facial expressions of pain and emotions," in *Forum Bildverarbeitung 2018*. KIT Scientific Publishing, 2018, p. 197.

[228] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening," *J. Sel. Topics Signal Processing*, vol. 4, no. 5, pp. 867–881, 2010.

[229] Y. Xu, N. Stojanovic, L. Stojanovic, and D. Kostic, "An approach for dynamic personal monitoring based on mobile complex event processing," in *Proceedings of International*

Conference on Advances in Mobile Computing & Multimedia, New York, NY, USA, 2013, pp. 464:464–464:473.

[230] C.-C. Yang and Y.-L. Hsu, "A review of accelerometry-based wearable motion detectors for physical activity monitoring," *Sensors*, vol. 10, no. 8, pp. 7772–7788, 2010.

[231] J. Yang, "Toward physical activity diary: Motion recognition using simple acceleration features with mobile phones," in *Proceedings of the 1st International Workshop on Interactive Multimedia for Consumer Electronics*, New York, NY, USA, 2009, pp. 1–10.

[232] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, January 2019.

[233] S. Yang and S. Cho, "Recognizing human activities from accelerometer and physiological sensors," in *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Aug 2008, pp. 100–105.

[234] T. Yang, G. Andrew, H. Eichner, H. Sun, W. Li, N. Kong, D. Ramage, and F. Beaufays, "Applied federated learning: Improving google keyboard query suggestions," *CoRR*, vol. abs/1812.02903, 2018.

[235] S. J. Young and S. Young, *The HTK hidden Markov model toolkit: Design and philosophy*. University of Cambridge, Department of Engineering Cambridge, England, 1993.

[236] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, "Audio-Visual Affective Expression Recognition Through Multistream Fused HMM," *Trans. Multi.*, vol. 10, no. 4, pp. 570–577, 2008.

[237] M. Zhang and A. A. Sawchuk, "Usc-had: A daily activity dataset for ubiquitous activity recognition using wearable sensors," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, New York, NY, USA, 2012, pp. 1036–1043.

[238] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Distributing recognition in computational paralinguistics," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 406–417, Oct 2014.

[239] C. Zimmermann, J. Zeilfelder, T. Bloecher, M. Diehl, S. Essig, and W. Stork, "Evaluation of a smart drink monitoring device," in *2017 IEEE Sensors Applications Symposium (SAS)*, March 2017, pp. 1–5.